Monotone finite difference discretization of degenerate elliptic partial differential equations using Voronoi's first reduction

Guillaume Bonnet

université PARIS-SACLAY

Monotone finite difference discretization of degenerate elliptic partial differential equations using Voronoi's first reduction

Discrétisation aux différences finies monotones d'équations aux dérivées partielles dégénérées elliptiques en utilisant la première réduction de Voronoi

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques appliquées
Unité de recherche : Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 1er décembre 2021, par

Guillaume BONNET

Composition du jury :

Quentin MÉRIGOT Professeur, Université Paris-Saclay	Président
Guillaume CARLIER Professeur, Université Paris-Dauphine – PSL et Inria Paris	Rapporteur & Examinateur
Ricardo NOCHETTO Professeur, University of Maryland, College Park	Rapporteur & Examinateur
Brittany FROESE Maîtresse de conférences (associate professor), NJIT	Examinatrice
Espen JAKOBSEN Professeur, NTNU	Examinateur

Direction de la thèse :

Jean-Marie MIREBEAU Directeur de recherche, CNRS et ENS Paris-Saclay, Université Paris-Saclay	Directeur
Frédéric BONNANS Directeur de recherche, Inria Saclay et CentraleSupélec, Université Paris-Saclay	Codirecteur

Invités :

Guy BARLES Professeur, Université de Tours	Invité
Xavier WARIN EDF R&D	Invité

Thèse de doctorat

NNT: 2021UPASM042


université
PARIS-SACLAY
FACULTÉ
DES SCIENCES
D'ORSAY


MATHEMATIQUES
ORSAY


CentraleSupélec
université
PARIS-SACLAY


L2S
Laboratoire
Signaux &
Systèmes


ÉCOLE
POLYTECHNIQUE
INSTITUT
POLYTECHNIQUE
DE PARIS


CMAP


Inria


Fondation mathématique
FMJH
Jacques Hadamard

Résumé

Dans cette thèse, nous concevons et étudions des discrétisations aux différences finies monotones sur grilles cartésiennes de certaines équations aux dérivées partielles dégénérées elliptiques. Ces discrétisations sont fondées sur la première réduction de Voronoi des formes quadratiques, un outil issu de l'étude de la géométrie des réseaux de petite dimension, ou sur l'algorithme de Selling, qui est une spécialisation de cet outil en dimensions deux et trois.

Dans la première partie de la thèse, nous étudions certaines propriétés de la première réduction de Voronoi et nous utilisons ces propriétés pour prouver certaines garanties théoriques à propos des schémas aux différences finis associés.

En dimensions deux et trois, nous recommandons une discrétisation particulière, consistante à l'ordre deux, d'opérateurs différentiels linéaires anisotropes comprenant à la fois des termes d'ordres un et deux. Nous prouvons que la construction recommandée est quasi-optimale, en termes de tailles de pas garantissant la monotonie de la discrétisation.

En dimensions allant jusqu'à quatre, nous étudions la régularité Lipschitz des coefficients et le rayon du stencil de discrétisations aux différences finies d'opérateurs de diffusion anisotropes construites en utilisant la première réduction de Voronoi, ainsi que l'absence d'artéfacts en damier dans les schémas qui en résultent.

Dans la seconde partie de la thèse, nous étudions des schémas aux différences finies monotones pour certaines équations aux dérivées partielles dégénérées elliptiques particulières.

Nous concevons une méthode numérique permettant d'approcher efficacement des distances de Randers. Cette méthode est fondée sur un principe de grandes déviations et se prête particulièrement bien à la résolution de la régularisation entropique de certains problèmes de transport optimal en utilisant l'algorithme de Sinkhorn. Nous montrons la convergence de la méthode. Nous étudions le choix optimal du paramètre survenant dans l'approximation par le principe de grandes déviations, par rapport au pas de discrétisation.

Nous discrétisons l'opérateur de Pucci bidimensionnel comme un maximum d'opérateurs linéaires discrets. Nous montrons que ce maximum admet une expression de forme fermée, ce qui réduit le coût numérique de son évaluation et donc aussi le coût de la résolution du schéma aux différences finies associé.

Nous discrétisons le second problème aux limites pour l'équation de Monge-Ampère. Nous prouvons l'existence de solutions à cette discrétisation. Nous prouvons aussi la convergence de ces solutions, lorsque l'équation de Monge-Ampère considérée est associée à un problème de transport optimal dont le coût est quadratique. En dimension deux, nous montrons que, similairement au cas de l'équation de Pucci, le maximum constituant l'opérateur de Monge-Ampère discrétisé admet une expression de forme fermée. Nous présentons une application numérique au problème du réfracteur en champ lointain en optique non imageante.

Abstract

In this thesis, we design and study monotone finite difference discretizations on Cartesian grids of some degenerate elliptic partial differential equations. These discretizations are based on Voronoi's first reduction of quadratic forms, a tool originating from the study of low-dimensional lattice geometry, or on Selling's algorithm, which is a specialization of this tool in dimensions two and three.

In the first part of the thesis, we study some properties of Voronoi's first reduction, and we use these properties to prove some theoretical guarantees about the associated finite difference schemes.

In dimensions two and three, we recommend a specific, second-order consistent discretization of linear anisotropic differential operators involving both a first- and a second-order term. We prove that the recommended construction is quasi-optimal, in terms of step sizes guaranteeing the monotonicity of the discretization.

In dimensions up to four, we study the Lipschitz regularity of the coefficients and the radius of the stencil of finite difference discretizations of anisotropic diffusion operators built using Voronoi's first reduction, as well as the absence of checkerboard artifacts in the resulting numerical schemes.

In the second part of the thesis, we study monotone finite difference schemes for some specific degenerate elliptic partial differential equations.

We design a numerical method allowing to efficiently approximate Randers distances. This method is based on a large deviations principle and lends itself particularly well to the resolution of the entropic regularization of some optimal transport problems using Sinkhorn's algorithm. We prove the convergence of the method. We study the optimal choice of the parameter occuring in the large deviations approximation with respect to the discretization step.

We discretize the two-dimensional Pucci operator as a maximum of linear discrete operators. We show that this maximum admits a closed-form formula, reducing the numerical cost of its evaluation and thus also the cost of solving the associated finite difference scheme.

We discretize the second boundary value problem for the Monge-Ampère equation. We prove the existence of solutions to this discretization. We also prove the convergence of these solutions, when the considered Monge-Ampère equation is associated to an optimal transport problem with a quadratic cost. In dimension two, we show that, similarly to the case of the Pucci equation, the maximum which constitutes the discretized Monge-Ampère operator admits a closed form formula. We present a numerical application to the far-field refractor problem in nonimaging optics.

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Jean-Marie Mirebeau et Frédéric Bonnans. Merci pour vos conseils, votre soutien et la confiance que vous m'avez accordée pendant ces trois ans. Dès le début de la thèse, vous avez su me proposer un projet de recherche motivant. Par la suite, votre enthousiasme permanent s'est manifesté à travers nos discussions fréquentes, qui ont été pour moi une importante source d'encouragement et au cours desquelles j'ai pu être inspiré par vos grandes curiosité et culture scientifiques.

I am very grateful to Guillaume Carlier and Ricardo Nochetto for the attention that they devoted to reviewing my manuscript and writing their reports. I want to thank Ricardo Nochetto for the helpful comments that he took the time to communicate to me about some specific points in this thesis. I thank Guy Barles, Brittany Froese, Espen Jakobsen, Quentin Mérigot, and Xavier Warin for having honored me by taking part in my examination committee.

J'ai préparé ma thèse dans un environnement de travail agréable et je souhaite remercier tous les membres du LMO, du CMAP et du L2S qui ont rendu cela possible, doctorants, postdoctorants, stagiaires ou permanents, chercheurs ou non, et plus particulièrement tous ceux avec qui j'ai eu l'occasion d'interagir lors des repas, des événements scientifiques ou sociaux et des tâches et rencontres diverses. J'adresse des remerciements spécifiques à la commission vie de laboratoire du CMAP, grâce à laquelle tous les doctorants du CMAP ont pu trouver leur place dans l'aile principale du laboratoire.

Ce qui est vrai à l'échelle du laboratoire l'est également à l'échelle du bureau. J'ai eu la chance de passer ces trois ans en bonne compagnie et je remercie vivement mes collègues de bureaux pour cela. Nos diverses discussions ont été pour moi autant d'occasions très appréciables de m'évader temporairement de mon pur domaine de recherche.

Merci aux coorganisateurs du CJC-MA 2021 pour avoir été et avoir su rester de très bons camarades pendant cette belle aventure de plus d'un an au cours de laquelle nous avons travaillé, communiqué et débattu avec une efficacité certes relative, mais que j'ai trouvée très satisfaisante de votre part compte tenu de notre faible expérience de la gestion de ce type de projet.

Je remercie enfin ma famille pour son soutien, de façon générale mais aussi et tout particulièrement, pour ce qui concerne ma famille proche, pendant la partie non négligeable de ces trois ans que j'ai eu, dans un contexte de crise sanitaire, l'occasion exceptionnelle de passer à vos côtés. Il a été pour moi aussi agréable qu'inattendu de pouvoir vous tenir compagnie pendant un peu plus longtemps que de simples vacances.

Contents

Chapitre 1

Introduction

Dans cette thèse, nous étudions la discrétisation aux différences finies monotones sur grille cartésienne d'équations aux dérivées partielles *dégénérées elliptiques*, c'est-à-dire, d'équations de la forme

$$F(x, u(x), Du(x), D^2u(x)) = 0,$$

où la fonction F est croissante par rapport à sa deuxième variable et décroissante par rapport à sa dernière variable. Nous nous concentrons principalement sur la discrétisation de l'opérateur de diffusion anisotrope sous forme non-divergence

$$u \mapsto \mathrm{Tr}(\mathcal{D}(\cdot)D^2u(\cdot)), \tag{1.1}$$

où $\mathcal{D}(\cdot)$ est un champ de matrices symétriques définies positives. Un exemple d'équation dégénérée elliptique comprenant un tel opérateur de diffusion est

$$\sup_{\alpha\in\mathcal{A}} \left(H^\alpha(x, u(x), Du(x)) - \mathrm{Tr}(\mathcal{D}^\alpha(x)D^2u(x))\right) = 0, \tag{1.2}$$

où $\mathcal{A}$ est un ensemble donné de paramètres, les fonctions H^α sont croissantes par rapport à leur deuxième variable et les $\mathcal{D}^\alpha$ sont des champs de matrices symétriques définies positives. Les équations de Pucci et de Monge-Ampère admettent des reformulations de la forme (1.2), voir les sections 1.2.2 et 1.2.3.

La propriété d'*ellipticité dégénérée* admet un équivalent discret, que nous appelons *monotonie* ou *ellipticité dégénérée discrète* selon la variante exacte de cet équivalent que nous considérons. Dans de nombreuses situations, un principe de comparaison discret entre les solutions de schémas numériques dégénérés elliptiques discrets peut être prouvé de façon très directe. De plus, un argument général a été introduit dans [BS91] pour prouver la convergence de schémas numériques monotones pour des équations aux dérivées partielles dégénérées elliptiques. Pour toutes ces raisons, il est souhaitable de concevoir des discrétisations d'équations dégénérées elliptiques qui satisfont la propriété de monotonie.

La discrétisation aux différences finies de l'opérateur (1.1) que nous considérons dans cette thèse est fondée sur une décomposition de toute matrice symétrique définie positive $\mathcal{D}$ sous la forme

$$\mathcal{D} = \sum_{i=1}^{I} \lambda_i e_i e_i^\top, \tag{1.3}$$

où les coefficients λ_i sont positifs ou nuls et les directions $e_i \in \mathbb{Z}^d$ sont des vecteurs à éléments entiers. Cette décomposition permet de construire l'approximation consistante à l'ordre deux

$$\mathrm{Tr}(\mathcal{D}D^2u(x)) \approx \sum_{i=1}^{I} \lambda_i \frac{u(x+he_i) + u(x-he_i) - 2u(x)}{h^2}. \tag{1.4}$$

Les éléments des vecteurs e_i doivent être entiers afin que seuls les points d'une grille cartésienne soient impliqués dans la discrétisation aux différences finies. Cette contrainte interdit d'utiliser simplement la décomposition en valeurs propres et vecteurs propres de la matrice $\mathcal{D}$. Les coefficients λ_i doivent être positifs ou nuls afin que la discrétisation soit utilisable dans des schémas numériques devant satisfaire la propriété de monotonie.

La stratégie décrite ci-dessus n'est pas le seul moyen de construire une discrétisation monotone et consistante de l'opérateur (1.1). Nous aurions pu choisir, par exemple, de relâcher la contrainte que les éléments des vecteurs e_i soient entiers et d'utiliser à la place une étape d'interpolation dans le schéma numérique pour approcher les valeurs de l'inconnue aux points qui n'appartiennent pas à la grille de discrétisation. Cette approche est connue [DJ13 ; NNZ19] et engendre une discrétisation à deux échelles de l'opérateur d'origine. Nous choisissons de maintenir la contrainte que les éléments des e_i soient entiers, en espérant obtenir des schémas aux différences finies avec des stencils de moindre rayon et des ordres de consistance plus élevés.

Historiquement, l'analyse numérique de schémas fondés sur une décomposition de la forme (1.3), satisfaisant les contraintes décrites ci-dessus, a été effectuée [Kry05 ; KT92], alors même qu'aucune méthode applicable numériquement n'était connue pour calculer cette décomposition, en dehors de certaines classes particulières de matrices $\mathcal{D}$ comme les matrices à diagonale dominante. Une construction fondée sur l'arbre de Stern-Brocot des fractions irréductibles a été introduite dans [BOZ04] pour des matrices de taille deux et une autre construction, équivalente pour des matrices de taille deux mais aussi applicable à des matrices de taille trois, et fondée sur l'algorithme de Selling [CS92 ; Sel74], un outil issu de l'étude de la géométrie des réseaux de petite dimension, a été introduite indépendamment dans [FM14]. La seconde construction possède une extension naturelle à des matrices symétriques définies positives de taille plus grande que trois, fondée sur la première réduction de Voronoi des formes quadratiques, qui est un outil connu dans la théorie de la géométrie des réseaux de petite dimension [CS88 ; Vor08] et a été utilisée précédemment pour la discrétisation d'équations aux dérivées partielles dans [Mir19].

La première partie de cette thèse est dévolue à l'étude de certaines propriétés de la première réduction de Voronoi qui présentent un intérêt lorsque celle-ci est appliquée à la conception de discrétisations aux différences finies. La seconde partie est dévolue à l'étude de schémas aux différences finies pour certaines équations dégénérées elliptiques particulières de la forme (1.2).

1.1 Outils pour la discrétisation aux différences finies monotones sur grille cartésienne d'opérateurs différentiels anisotropes

Pour une matrice symétrique définie positive $\mathcal{D}$ donnée, nous recommandons de choisir les coefficients de sa décomposition (1.3) comme une solution du problème de maximisation

$$\max \left\{ \sum_{e \in \mathbb{Z}^d \setminus \{0\}} \lambda^e \;\middle|\; \lambda \colon \mathbb{Z}^d \setminus \{0\} \to \mathbb{R}_+, \sum_{e \in \mathbb{Z}^d \setminus \{0\}} \lambda^e e e^\top = \mathcal{D} \right\}. \tag{1.5}$$

Bien que ce problème soit posé sur l'ensemble de toutes les décompositions de $\mathcal{D}$, à support potentiellement infini, dont les coefficients sont positifs ou nuls et dont les vecteurs sont à éléments entiers, il peut être montré que ses solutions sont toujours à support fini. De plus, puisque la contrainte $\sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e ee^\top = \mathcal{D}$ implique que $\sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e |e|^2 = \mathrm{Tr}(\mathcal{D})$, le fait de maximiser les coefficients λ^e est cohérent avec le fait qu'il soit souhaitable, pour des raisons numériques, que les vecteurs $e \in \mathbb{Z}^d \setminus \{0\}$ associés aux coefficients λ^e non nuls soient de norme petite.

Le problème de maximisation (1.5) admet le problème dual suivant :

$$\min_{M\in\mathcal{M}_d} \mathrm{Tr}(\mathcal{D}M); \quad \mathcal{M}_d := \{M \in \mathbb{R}^{d\times d} \text{ symétrique}, \langle e, Me\rangle \geq 1, \forall e \in \mathbb{Z}^d \setminus \{0\}\}. \tag{1.6}$$

Il se trouve que l'ensemble admissible $\mathcal{M}_d$ de ce problème dual est connu dans le domaine de la géométrie des réseaux de petite dimension, sous le nom de *polyèdre de Ryskov*. Bien qu'il soit défini par une infinité de contraintes, il présente localement la même structure qu'un polyèdre standard [Sch09a].

Afin d'expliquer le lien entre le polyèdre de Ryskov et la géométrie des réseaux de petite dimension, associons à toute matrice symétrique définie positive M de taille d le réseau, unique à rotation près, engendré par les combinaisons linéaires à coefficients entiers des colonnes d'une matrice B de taille d telle que $B^\top B = M$. Alors le polyèdre de Ryskov est l'ensemble de toutes les matrices symétriques définies positives associées à des réseaux pour lesquels un empilement de sphères ne se chevauchant pas peut être construit en plaçant une sphère de rayon un demi centrée en chaque point du réseau, comme illustré par la Figure 1.1. Les sommets du polyèdre de Ryskov sont appelés des *formes parfaites* et les réseaux associés sont appelés des *réseaux parfaits*.

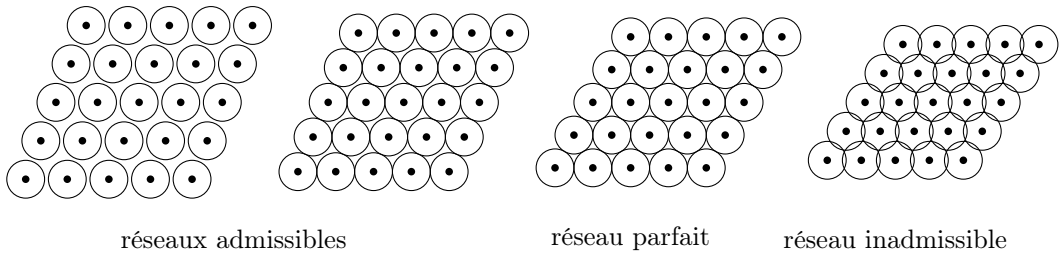


FIGURE 1.1 : Polyèdre de Ryskov : interprétation en termes de réseaux. Dans le cas limite d'un réseau parfait, chaque sphère est en contact avec au moins $d(d+1)$ sphères voisines (c'est-à-dire 6 sphères voisines en dimension $d = 2$).

Bien que, dans les applications à l'étude des empilements de sphères, la quantité minimisée sur le polyèdre de Ryskov soit habituellement le déterminant, plutôt qu'une forme linéaire comme dans le problème (1.6), les deux problèmes de minimisation partagent la propriété que leur minimum soit atteint en une forme parfaite, c'est-à-dire en un sommet du polyèdre de Ryskov. Cela a motivé la classification des formes parfaites, qui a été effectuée dans la littérature jusqu'à la dimension $d = 8$ [CS88 ; DSV07]. La classification des matrices symétriques définies positives $\mathcal{D}$ en fonction de quelle forme parfaite M est optimale dans (1.6) est appelée la *première réduction de Voronoi des formes quadratiques*. Deux formes parfaites M et M' sont dites *arithmétiquement équivalentes* si elles sont associées au même réseau, ou bien, de façon équivalente, s'il existe une matrice A de taille d à éléments entiers et dont l'inverse est également à éléments entiers telle que $M' = A^\top MA$. Bien qu'il y ait, en toute dimension $d \geq 2$, une infinité de formes parfaites, Voronoi a prouvé qu'il existe seulement un nombre fini de classes d'équivalence de formes parfaites pour la relation d'équivalence arithmétique [Vor08]. Le fait de connaître ces classes d'équivalence, ainsi que la façon dont la relation d'équivalence arithmétique interagit avec la structure polyédrale du

polyèdre de Ryskov, permet de résoudre particulièrement efficacement le problème dual (1.6) — et donc aussi le problème primal (1.5) — en parcourant le graphe d'adjacence des formes parfaites.

Remarquez que le nombre de classes d'équivalence de formes parfaites explose avec la dimension, comme illustré par la Table 1.1. C'est pourquoi l'approche décrite dans cette thèse pour construire des décompositions de la forme (1.3) est applicable en pratique pourvu que les dimensions considérées restent raisonnablement petites.

$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$	$d\geq 9$
1	1	2	3	7	33	10916	inconnu

TABLE 1.1 : Nombre de classes d'équivalence de formes parfaites en fonction de la dimension.

En dimensions $d \in \{2, 3\}$, la structure de cette classification est particulièrement simple, puisqu'il existe uniquement une classe d'équivalence de formes parfaites pour la relation d'équivalence arithmétique. Le problème de minimisation (1.6) peut donc être résolu en utilisant des outils spécifiques à ces dimensions, comme l'algorithme de Selling [CS92 ; Sel74], qui a été utilisé précédemment pour la discrétisation d'équations aux dérivées partielles [FM14]. Dans le Chapitre 2, nous montrons comment voir la décomposition (1.3) obtenue avec l'algorithme de Selling comme une solution du problème de maximisation (1.5) permet de prouver certaines garanties théoriques à propos de la faisabilité de la discrétisation monotone et consistante à l'ordre deux d'un opérateur différentiel comprenant à la fois des termes d'ordres un et deux.

Dans le Chapitre 3, nous expliquons comment l'ensemble de solutions du problème de maximisation (1.5) peut être calculé efficacement en dimension $d = 4$, nous recommandons un moyen de choisir une solution particulière lorsque cet ensemble n'est pas un singleton et nous discutons certaines propriétés des schémas aux différences finies qui en résultent, à savoir la régularité Lipschitz de leurs coefficients, le rayon de leurs stencils et l'absence d'artéfacts en damier.

1.1.1 Discrétisation aux différences finies monotones d'ordre deux d'opérateurs différentiels linéaires anisotropes

En dimensions $d \in \{2, 3\}$, toutes les formes parfaites sont équivalentes, pour la relation d'équivalence arithmétique discutée ci-dessus, à la forme parfaite de référence

$$\frac{1}{2}\begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 2 \end{pmatrix} = \frac{1}{2}I_d + \frac{1}{2}\mathbb{1}\mathbb{1}^\top. \tag{1.7}$$

Pour cette raison, toutes les décompositions de la forme (1.3), obtenues en utilisant la première réduction de Voronoi des formes quadratiques, de toutes les matrices symétriques définies positives $\mathcal{D}$ de taille deux ou trois ont des ensembles de vecteurs directions partageant la même structure particulière, liée à la notion de superbase du réseau $\mathbb{Z}^d$ (une base de $\mathbb{Z}^d$ est une famille $v = (v_1, \dots, v_d)$ de vecteurs à éléments entiers satisfaisant $\det(v_1, \dots, v_d) = \pm 1$; une superbase de $\mathbb{Z}^d$ est une base de $\mathbb{Z}^d$ étendue par un vecteur additionnel $v_0 := -v_1 - \dots - v_d$). Plus précisément, il existe une superbase v de $\mathbb{Z}^d$ telle que la décomposition est supportée par les vecteurs $\pm e_{ij}$, $i,\ j \in \{0, \dots, d\}$, définis par $\pm\langle e_{ij}, v_k\rangle = \pm(\delta_{ik} - \delta_{jk})$, pour tout $k \in \{0, \dots, d\}$. De plus, le coefficient $\lambda^{e_{ij}}$ associé au vecteur e_{ij} est égal au produit scalaire $-\langle v_i, \mathcal{D}v_j\rangle$ et la contrainte qu'il doive être positif ou nul peut s'interpréter comme une propriété d'*angle obtus* de la superbase v pour le produit scalaire concerné.

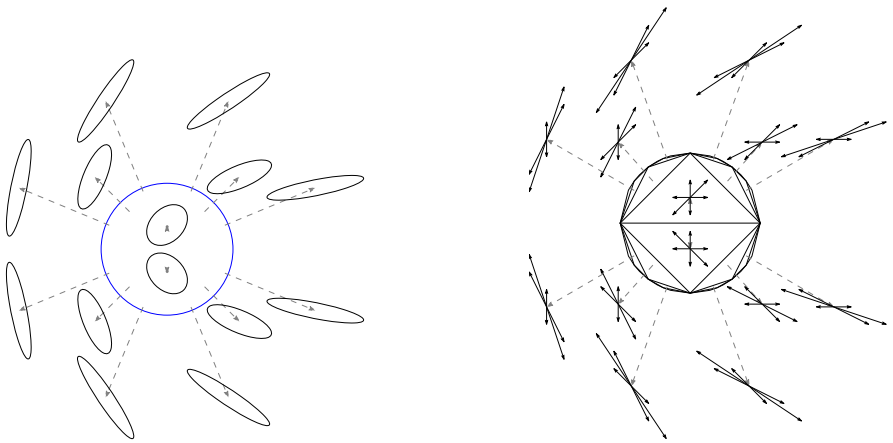


FIGURE 1.2 : À gauche : l'ensemble des matrices symétriques définies positives de taille deux et de trace un. À droite : les stencils de différences finies associés.

Nous affichons dans la Figure 1.2 les ensembles de vecteurs directions dans les décompositions de certaines matrices symétriques définies positives de taille deux. En dimension $d = 2$, l'ensemble des matrices symétriques définies positives dont la trace a été normalisée à un est un disque, qui peut être paramétrisé comme

$$\left\{ \frac{1}{2} \begin{pmatrix} 1+\rho_1 & \rho_2 \\ \rho_2 & 1-\rho_1 \end{pmatrix} \;\middle|\; \rho_1^2 + \rho_2^2 \leq 1 \right\}.$$

Les vecteurs directions sont constants sur des triangles, qui forment une triangulation infinie du disque et qui coïncident avec les cellules de la première réduction de Voronoi de l'ensemble des formes quadratiques de dimension deux de trace normalisée. En dimension deux, les vecteurs directions de la décomposition coïncident eux-mêmes, à un changement de signe près, avec les éléments d'une superbase de $\mathbb{Z}^2$.

Alors que la discrétisation aux différences finies d'opérateurs différentiels anisotropes du second ordre en dimensions $d \in \{2, 3\}$, fondée sur une décomposition de la matrice coefficient obtenue en utilisant la structure particulière des formes parfaites en dimensions $d \in \{2, 3\}$, a été étudiée précédemment [FM14], nous considérons dans le Chapitre 2 des opérateurs différentiels comprenant à la fois des termes du premier et du second ordres. Il existe au moins deux stratégies de discrétisations usuelles pour le terme du premier ordre : celle utilisant des différences finies décentrées amont, qui sont consistantes à l'ordre un et monotones, et celle utilisant des différences finies centrées, qui sont consistantes à l'ordre deux mais auxquelles la monotonie fait défaut. En dimension un, il est bien connu que, en présence d'un terme du second ordre non dégénéré, le défaut de monotonie de la discrétisation aux différences finies centrées du terme du premier ordre peut être compensé par la monotonie de la discrétisation du terme du second ordre, permettant de discrétiser l'opérateur complet de façon consistante à l'ordre deux. Nous étudions comment cette construction peut être étendue à des opérateurs différentiels de dimensions deux et trois, de la forme

$$u \mapsto \langle b(\cdot), Du(\cdot) \rangle + \mathrm{Tr}(\mathcal{D}(\cdot) D^2 u(\cdot)),$$

où b et $\mathcal{D}$ sont des champs respectivement de vecteurs et de matrices symétriques définies positives.

Nous introduisons la discrétisation consistante à l'ordre deux

$$\begin{aligned}\langle b, Du(x)\rangle + \mathrm{Tr}(\mathcal{D}D^2u(x)) \approx &\sum_{i=1}^{I} \mu_i \frac{u(x+he_i) - u(x-he_i)}{2h} \\ &+ \sum_{i=1}^{I} \lambda_i \frac{u(x+he_i) + u(x-he_i) - 2u(x)}{h^2},\end{aligned} \tag{1.8}$$

où les λ_i et les e_i sont respectivement les coefficients et les vecteurs d'une décomposition de la matrice symétrique définie positive $\mathcal{D}$ sous la forme (1.3) et où les μ_i sont les coefficients d'une décomposition du vecteur b sous la forme

$$b = \sum_{i=1}^{I} \mu_i e_i. \tag{1.9}$$

Il est important que les mêmes vecteurs e_i soient partagés entre les décompositions de $\mathcal{D}$ et de b, puisque sinon le défaut de monotonie de la discrétisation du terme du premier ordre ne pourrait pas être compensée.

Nous recommandons de choisir la décomposition (1.3) fondée sur la première réduction de Voronoi des formes quadratiques, après l'avoir calculée en utilisant l'algorithme de Selling, et de choisir ensuite les coefficients μ_i d'après la formule

$$\mu_i := \lambda_i \langle b, \mathcal{D}^{-1} e_i\rangle.$$

Nous montrons (Théorème 2.1.7) qu'avec ce choix de coefficients et de vecteurs, la discrétisation (1.8) est monotone pourvu que le pas de discrétisation h soit suffisamment petit, plus précisément plus petit que $C|\mathcal{D}^{-1}|^{-1/2}\langle b, \mathcal{D}^{-1}b\rangle^{-1/2}$, où la constante C dépend seulement de la dimension. Le principal résultat du Chapitre 2 est que ce choix est quasi-optimal, en ce sens que si la discrétisation (1.8) est monotone pour un certain choix de coefficients et de vecteurs, alors elle est aussi monotone pour le choix de coefficients et de vecteurs que nous proposons, à une division près du pas de discrétisation h par un facteur deux en dimension deux ou par un facteur six en dimension trois (Théorème 2.1.6).

Afin de prouver le résultat principal, nous devons étudier une extension de la première réduction de Voronoi des formes quadratiques aux formes quadratiques inhomogènes. Cette extension implique la variante suivante du polyèdre de Ryskov :

$$\widetilde{\mathcal{M}}_d := \{(\eta, M) \mid \eta \in \mathbb{R}^d,\, M \in \mathbb{R}^{d\times d} \text{ symétrique};\, \forall e \in \mathbb{Z}^d \setminus \{0\},\, \langle \eta, e\rangle + \langle e, Me\rangle \geq 1\}.$$

Alors que $\mathcal{M}_d$, la variante usuelle du polyèdre de Ryskov, est un ensemble de matrices symétriques, le polyèdre $\widetilde{\mathcal{M}}_d$ que nous introduisons est un ensemble de paires de vecteurs et de matrices symétriques. Nous montrons que la structure polyédrale de $\widetilde{\mathcal{M}}_d$ est remarquablement similaire à celle du polyèdre de Ryskov $\mathcal{M}_d$: par exemple, tous les sommets de $\widetilde{\mathcal{M}}_d$ sont de la forme $(0, M)$, où M est un sommet de $\mathcal{M}_d$.

1.1.2 Discrétisation monotone d'opérateurs différentiels anisotropes en dimension quatre en utilisant la première réduction de Voronoi

En dimension $d = 4$, à la différence des dimensions deux et trois, les formes parfaites ne sont pas toutes arithmétiquement équivalentes à la forme parfaite de référence (1.7) et il existe exactement

une autre classe d'équivalence de formes parfaites, dont un représentant est

$$\frac{1}{2}\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}. \tag{1.10}$$

Les formes parfaites arithmétiquement équivalentes à (1.10) ne sont pas liées à des *superbases* de $\mathbb{Z}^d$ comme celles arithmétiquement équivalentes à (1.7) le sont. C'est pourquoi l'algorithme de Selling, qui consiste en une itération sur les superbases de $\mathbb{Z}^d$, n'est pas directement applicable en dimension $d = 4$. Cependant, le problème de minimisation (1.6) peut toujours être résolu en parcourant les sommets du polyèdre de Ryskov de façon appropriée. Nous expliquons dans le Chapitre 3 que cette procédure peut être implémentée particulièrement efficacement, en utilisant le fait que les relations d'adjacence entre les formes parfaites, dans le graphe des sommets du polyèdre de Ryskov, peuvent être précalculées.

d	forme parfaite arithmétiquement équivalente à	nombre de voisines arithmétiquement équivalentes à		dimension de l'espace des matrices symétriques de taille d	forme parfaite dégénérée
		(1.7)	(1.10)		
2	(1.7)	3	0	3	non
3	(1.7)	6	0	6	non
4	(1.7)	0	10	10	non
4	(1.10)	48	16	10	oui

TABLE 1.2 : Structure polyédrale du polyèdre de Ryskov en dimensions $d \in \{2, 3, 4\}$.

Nous décrivons comment l'ensemble des solutions du problème de maximisation (1.5) peut en être déduit. Cet ensemble n'est pas toujours un singleton. Ceci est lié au fait que, comme illustré par la Table 1.2, certaines formes parfaites — celles qui sont arithmétiquement équivalentes à (1.10) — sont des sommets dégénérés du polyèdre de Ryskov, c'est-à-dire qu'elles ont plus de dix voisines, dix étant la dimension de l'espace des matrices symétriques de taille quatre. Nous montrons cependant que, en dimension $d = 4$ et à identification près des vecteurs $e \in \mathbb{Z}^4 \setminus \{0\}$ avec leurs opposés, l'ensemble des solutions de (1.5) est toujours un triangle équilatéral, potentiellement réduit à un singleton. Nous recommandons de choisir le barycentre de cet ensemble comme la décomposition de la matrice symétrique définie positive $\mathcal{D}$ à utiliser dans des discrétisations aux différences finies et nous montrons quelques propriétés des discrétisations qui en résultent :

Régularité Lipschitz des coefficients. Nous montrons (Théorème 3.3.6) que les coefficients λ^e de la décomposition proposée de la matrice symétrique définie positive $\mathcal{D}$ dépendent de $\mathcal{D}$ de façon localement Lipschitz, la constante de Lipschitz dépendant seulement de conditionnement de $\mathcal{D}$. Comme corollaire, lors de la discrétisation d'un opérateur de diffusion anisotrope tel que (1.1) impliquant un champ de matrices symétriques définies positives $\mathcal{D}(\cdot)$ qui est localement Lipschitz et dont le conditionnement est borné, les champs de coefficients $\lambda^e(\cdot)$ de la discrétisation restent Lipschitz. Un des avantages de prouver la continuité des coefficients de la discrétisation est qu'il s'agit d'une hypothèse de certains résultats connus à propos des vitesses de convergence de certains schémas numériques pour certaines équations aux dérivées partielles [BJ07].

Rayon du stencil de différences finies. Le stencil de la discrétisation aux différences finies d'un opérateur de diffusion anisotrope associée à la décomposition d'une matrice symétrique définie positive $\mathcal{D}$ qui est solution du problème de maximisation (1.5) est l'ensemble des vecteurs

$e \in \mathbb{Z}^d \setminus \{0\}$ associés aux coefficients λ^e non nuls. Pour une meilleure efficacité du schéma aux différences finies, il est souhaitable que ces vecteurs soient de norme petite. Nous prouvons l'estimation $|e| \leq C\mu(\mathcal{D})$, où C est une constante dépendant seulement de la dimension et $\mu(\mathcal{D}) := |\mathcal{D}|^{1/2}|\mathcal{D}^{-1}|^{1/2}$ est la racine carrée du conditionnement de la matrice $\mathcal{D}$ (Théorème 3.4.1). La valeur de la constante C est importante et est discutée dans la section 3.4.1.

L'estimation ci-dessus est vraie en toute dimension $d \in \mathbb{N}^*$ et pour des décompositions définies par tout λ maximal dans (1.5). Elle a été prouvée précédemment en dimensions $d \in \{2, 3\}$, alors qu'en dimensions supérieures seule l'estimation plus faible $|e| \leq C\mu(\mathcal{D})^{d-1}$ était connue, voir [Mir19]. Afin de prouver l'estimation améliorée, nous devons répondre à la question suivante : sachant que $\mathcal{D}$ appartient à une cellule particulière de la première réduction de Voronoi des formes quadratiques, existe-t-il un nombre fini de cellules de cette réduction à l'union desquelles la matrice $\mathcal{D}^{-1}$ est garantie d'appartenir ? Nous montrons que cela est vrai, en utilisant que les cellules de la première réduction de Voronoi sont des enveloppes coniques convexes d'ensembles de matrices de rang un, ce qui simplifie grandement la structure de leur image par la fonction inverse matricielle.

Garanties contre les artéfacts en damier. Les artéfacts en damier surviennent habituellement lorsque le graphe d'adjacence des points d'une grille de discrétisation, conformément aux stencils d'un schéma aux différences finies, n'est pas connexe, comme illustré par la Figure 1.3. Alors les restrictions du schéma aux différentes composantes connexes de la grille se comportent de façon indépendante, ce qui peut être malvenu. Nous présentons une stratégie permettant de prouver l'absence de tels artéfacts dans certaines discrétisations d'opérateurs de dimension quatre comprenant une diffusion anisotrope, en utilisant les propriétés précédemment prouvées sur le caractère Lipschitz des coefficients λ^e et sur la norme des vecteurs e de la décomposition recommandée d'une matrice symétrique définie positive $\mathcal{D}$, ainsi que la *propriété d'engendrement* suivante : l'ensemble des vecteurs e associés à des coefficients λ^e non nuls génère le réseau $\mathbb{Z}^4$ par combinaisons linéaires à coefficients entiers. Nous montrons (Théorème 3.5.1) qu'alors qu'il existe des décompositions de $\mathcal{D}$ associées à des maximiseurs λ du problème (1.5) qui ne satisfont pas cette dernière propriété, la décomposition recommandée, qui est associée au barycentre de l'ensemble des maximiseurs, la satisfait.

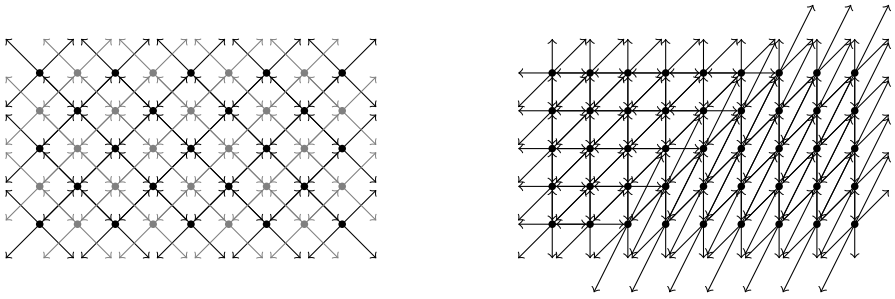


schéma avec des artéfacts en damier — schéma sans artéfacts en damier

FIGURE 1.3 : À gauche : stencils d'une discrétisation de l'équation de Laplace utilisant des directions diagonales, qui est un exemple de schéma entraînant des artéfacts en damier. À droite : stencils associés à la discrétisation recommandée dans cette thèse pour un opérateur de diffusion anisotrope, possédant de bonnes propriétés de connexité qui empêchent les artéfacts en damier.

1.2 Discrétisation monotone de quelques équations aux dérivées partielles dégénérées elliptiques particulières

Dans la seconde partie de cette thèse, nous étudions certains schémas aux différences finies particuliers, conçus en utilisant la stratégie précédemment discutée pour la discrétisation d'opérateurs différentiels anisotropes. Ces schémas nous permettent d'une part d'approcher des distances de Randers et des distances de transport optimal associées et d'autre part de résoudre numériquement les équations de Pucci et de Monge-Ampère, reformulées sous la forme (1.2).

1.2.1 Un schéma aux différences finies linéaire pour approcher la distance de Randers sur une grille cartésienne

Distances de Randers. Les distances de Randers sont une extension asymétrique des distances riemanniennes. Leur asymétrie peut illustrer, par exemple, le fait qu'il soit plus facile de se déplacer dans le sens du courant que dans le sens opposé, dans un milieu sujet à des courants, ou plus facile de descendre une pente que de la monter à cause de l'effet de la gravité. Le problème de navigation de Zermelo [BRS04], illustré dans la Figure 1.4, est un exemple de problème connu impliquant des distances de Randers.

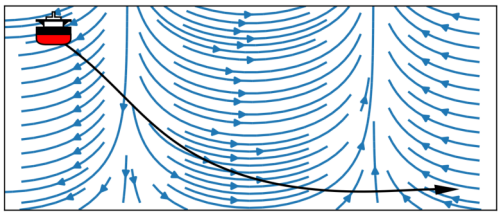


FIGURE 1.4 : Le problème de navigation de Zermelo. L'objectif est de calculer la trajectoire (en noir) permettant au navire de rejoindre sa destination en un temps minimal, en prenant en compte les courants marins (en bleu). Le temps de trajet minimal entre une source et une destination données définit une distance de Randers.

Les distances de Randers ont initialement été introduites dans le cadre de la relativité générale [Ran41]. Elles ont de nombreuses autres applications, dont la segmentation d'images [CMC16], les cortex quantiques [ABM06] et la pénalisation de courbure de chemins [CMC17].

Une métrique de Randers dans un domaine ouvert $\Omega \subset \mathbb{R}^d$ est une fonction $\mathcal{F}\colon \overline{\Omega} \times \mathbb{R}^d \to \mathbb{R}$ de la forme

$$\mathcal{F}_x(v) := \langle v, M(x)v\rangle^{1/2} + \langle \omega(x), v\rangle,$$

où M et ω sont des champs donnés, respectivement de matrices symétriques définies positives et de vecteurs, dans Ω. La condition de compatibilité $\langle \omega(x), M(x)^{-1}\omega(x)\rangle < 1$ entre les deux champs est supposée dans Ω. Dans le cas particulier d'un champ de vecteurs ω identiquement nul, la métrique de Randers est réduite à une métrique riemannienne.

À deux points x, $y \in \overline{\Omega}$ est associé l'ensemble Γ_x^y de tous les chemins Lipschitz $\gamma\colon [0,1] \to \overline{\Omega}$ entre x et y (c'est-à-dire satisfaisant $\gamma(0) = x$ et $\gamma(1) = y$). La longueur d'un tel chemin par rapport à une métrique de Randers $\mathcal{F}$ donnée est définie comme suit :

$$\text{longueur}_{\mathcal{F}}(\gamma) := \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t))\, dt.$$

La distance de Randers entre les points x et y est définie comme la longueur minimale parmi tous les chemins :

$$\mathrm{dist}_{\mathcal{F}}(x,y) := \inf_{\gamma\in\Gamma_x^y} \mathrm{longueur}_{\mathcal{F}}(\gamma).$$

Dans le Chapitre 4, nous introduisons une méthode numérique pour approcher la fonction $\mathbf{u}$ définie dans $\overline{\Omega}$ par

$$\mathbf{u}(x) := \inf_{p\in\partial\Omega} (g(p) + \mathrm{dist}_{\mathcal{F}}(p,x)), \tag{1.11}$$

pour une fonction $g\colon \partial\Omega \to \mathbb{R}\cup\{+\infty\}$ donnée (cela inclut le cas particulier d'une fonction $\mathbf{u}\colon x \mapsto \mathrm{dist}_{\mathcal{F}}(p_0,x)$, quitte à exclure le point p_0 de Ω et à choisir $g(p_0)=0$ et $g=+\infty$ sur $\partial\Omega\setminus\{p_0\}$). Cette méthode implique la résolution d'un schéma aux différences finies linéaire et est justifiée par un principe de grandes déviations. Ses avantages incluent d'une part que la linéarité du schéma permet d'utiliser des techniques de préfactorisation pour approcher efficacement des distances entre de nombreuses paires de points et d'autre part qu'elle est particulièrement appropriée à la résolution numérique de la régularisation entropique de problèmes de transport optimal. Une méthode similaire a précédemment été introduite dans le cadre particulier des variétés riemanniennes [CWW13] et a été appliquée à l'approximation de distances de transport optimal dans de telles variétés [Sol+15].

Le schéma aux différences finies que nous recommandons de résoudre est de la forme

$$\begin{aligned} &u_\varepsilon^h(x) + 2\varepsilon \sum_{i=1}^{I} \mu_i(x) \frac{u(x+he_i(x)) - u(x-he_i(x))}{2h} \\ &- \varepsilon^2 \sum_{i=1}^{I} \lambda_i(x) \frac{u(x+he_i(x)) + u(x-he_i(x)) - 2u(x)}{h^2} = 0 \quad \text{dans } \Omega\cap h\mathbb{Z}^d, \end{aligned} \tag{1.12}$$

où les coefficients μ_i et λ_i et les vecteurs e_i satisfont (1.3) et (1.9), b et $\mathcal{D}$ étant des champs dans Ω, respectivement de vecteurs et de matrices symétriques définies positives, définis à partir de ω et M par des relations algébriques simples. Remarquez la similarité entre le membre de gauche dans le schéma (1.12) et la discrétisation (1.8). Le schéma doit être adapté près de $\partial\Omega$ pour prendre en compte la condition aux limites de Dirichlet $u_\varepsilon^h(x) = \exp(-g(x)/\varepsilon)$ sur $\partial\Omega$. Nous affirmons que la fonction $\mathbf{u}$ est approchée, sous des hypothèses appropriées, par $\mathbf{u}_\varepsilon^h := -\varepsilon \ln u_\varepsilon^h$, pour des petites valeurs de ε et de h.

Nous affichons dans la Figure 1.5 des résultats numériques obtenus en appliquant la méthode numérique proposée au problème de navigation de Zermelo.

Principe de grandes déviations. Il est bien connu que la fonction $\mathbf{u}$ définie par (1.11) est solution de l'équation de Hamilton-Jacobi-Bellman

$$\langle D\mathbf{u}(x), \mathcal{D}(x) D\mathbf{u}(x)\rangle + 2\langle b(x), D\mathbf{u}(x)\rangle - 1 = 0 \quad \text{dans } \Omega, \tag{1.13}$$

où $\mathcal{D}$ et b sont les champs introduits ci-dessus. D'autre part, le schéma (1.12) est une discrétisation de l'équation linéaire d'ordre deux

$$u_\varepsilon(x) + 2\varepsilon\langle b(x), Du_\varepsilon(x)\rangle - \varepsilon^2 \operatorname{Tr}(\mathcal{D}(x) D^2 u_\varepsilon(x)) = 0 \quad \text{dans } \Omega. \tag{1.14}$$

Formellement, il est facile de montrer que si u_ε est solution de (1.14), alors $\mathbf{u}_\varepsilon := -\varepsilon \ln u_\varepsilon$ est solution de

$$\langle D\mathbf{u}_\varepsilon(x), \mathcal{D}(x) D\mathbf{u}_\varepsilon(x)\rangle + 2\langle b(x), D\mathbf{u}_\varepsilon(x)\rangle - \varepsilon \operatorname{Tr}(\mathcal{D}(x) D^2\mathbf{u}_\varepsilon(x)) - 1 = 0 \quad \text{dans } \Omega. \tag{1.15}$$

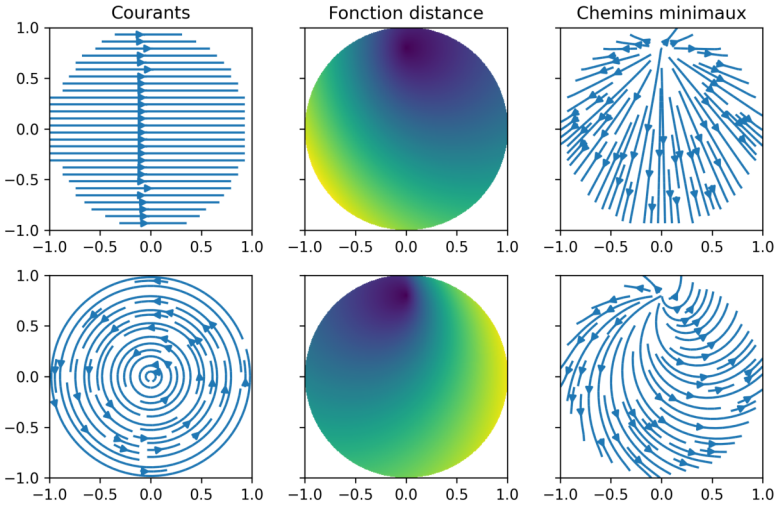


FIGURE 1.5 : Application de la méthode numérique proposée au problème de navigation de Zermelo (voir Figure 1.4). À gauche : courants ambiants. Au milieu : fonction distance de Randers depuis le point $p_0 := (0, 0.8)$. À droite : chemins minimaux depuis le point p_0.

L'équation de Hamilton-Jacobi-Bellman (1.15) est une perturbation de (1.13), ce qui justifie la convergence de $\mathbf{u}_\varepsilon$ vers $\mathbf{u}$ lorsque ε tend vers zéro.

La justification ci-dessus est formelle ; nous expliquons dans la section 4.A comment elle peut être rendue rigoureuse dans le cadre des solutions de viscosité.

Nous qualifions le résultat $\mathbf{u}_\varepsilon \to_{\varepsilon\to 0} \mathbf{u}$ de principe de grandes déviations parce que la fonction $\mathbf{u}_\varepsilon$ a l'interprétation probabiliste suivante : pour $x \in \overline{\Omega}$ et $\varepsilon > 0$ donnés, soit $(X_t^{x,\varepsilon})_{t\geq 0}$ le processus stochastique défini par

$$dX_t^{x,\varepsilon} = -2\varepsilon b(X_t^{x,\varepsilon})\,dt + \varepsilon\sqrt{2\mathcal{D}}(X_t^{x,\varepsilon})\,dW_t, \qquad X_0^{x,\varepsilon} = x,$$

où $(W_t)_{t\geq 0}$ est un mouvement d-dimensionnel et soit $\tau^{x,\varepsilon} \geq 0$ le temps de sortie

$$\tau^{x,\varepsilon} := \inf\{t \geq 0 \mid X_t^{x,\varepsilon} \notin \Omega\}.$$

Alors $\mathbf{u}_\varepsilon(x)$ est égal à

$$-\varepsilon \ln\left(\mathbb{E}\left[\exp\left(\frac{-\varepsilon\tau^{x,\varepsilon} - g(X_{\tau^{x,\varepsilon}}^{x,\varepsilon})}{\varepsilon}\right)\right]\right),$$

qui peut s'interpréter comme un soft-minimum et se comparer à l'infimum dans (1.11).

Analyse de convergence. En plus de la convergence $\mathbf{u}_\varepsilon \to_{\varepsilon\to 0} \mathbf{u}$, il est facilement prouvé, sous des hypothèses appropriées, en utilisant la théorie usuelle de la convergence des schémas aux différences finies pour des équations linéaires et le fait que le schéma (1.12) soit consistent avec l'équation (1.14), que $u_\varepsilon^h \to_{h\to 0} u_\varepsilon$, et donc aussi que $\mathbf{u}_\varepsilon^h \to_{h\to 0} \mathbf{u}_\varepsilon$. Cependant, cela ne garantit pas la convergence jointe $\mathbf{u}_\varepsilon^h \to_{(\varepsilon,h)\to 0} \mathbf{u}$.

Nous étudions la convergence jointe dans deux cadres différents. Dans le premier, que nous appelons le *cadre régulier*, nous supposons, entre autres hypothèses, que le domaine Ω est régulier et que la fonction g dans (1.11) est continue et prend des valeurs finies. Dans le second, que nous

appelons le *cas singulier*, nous supposons que le bord de Ω est l'union d'une partie régulière et d'un point isolé p_0 et nous choisissons $g(p_0) = 0$ et $g = +\infty$ ailleurs : comme déjà discuté plus haut, cela nous permet de considérer le cas de la fonction $\mathbf{u}\colon x \mapsto \mathrm{dist}_{\mathcal{F}}(p_0, x)$, qui est souvent celle qui doit être approchée dans les applications pratiques.

Dans le cas régulier, notre stratégie est, à la place d'effectuer une transformation logarithmique dans l'équation linéaire (1.14) afin d'obtenir l'équation non linéaire (1.15), de plutôt effectuer la transformation logarithmique directement dans le schéma aux différences finies linéaire (1.12), ce qui engendre un schéma non linéaire et monotone dont $\mathbf{u}_\varepsilon^h := -\varepsilon \ln u_\varepsilon^h$ est solution. Sous des hypothèses appropriées, nous montrons (Proposition 4.3.13) la consistance de ce schéma non linéaire avec l'équation (1.13) lorsque $\varepsilon \to 0$ et $h/\varepsilon \to 0$ et nous en déduisons (Théorème 4.3.18) la convergence de $\mathbf{u}_\varepsilon^h$ vers $\mathbf{u}$. De façon importante, le résultat de convergence n'est plus valide lorsque le rapport h/ε reste constant, ce qui rappelle le contre-exemple à la convergence de la méthode décrite dans [CWW13] vers la fonction distance, voir [CWW13, Appendice A]. Nous recommandons de choisir le paramètre ε proportionnellement à $h^{2/3}$. Nous montrons (Corollaire 4.3.14) que, au moins loin de $\partial\Omega$, le schéma non linéaire est consistant à l'ordre $2/3$ dans ce cas, ce qui est le meilleur ordre de consistance parmi tous les choix possibles du paramètre ε. À titre de comparaison, nous montrons que si des différences finies décentrées amont avaient été utilisées à la place de différences finies centrées pour la discrétisation du terme du premier ordre dans le schéma (1.12), alors le schéma non linéaire aurait seulement été consistant à l'ordre $1/2$.

Dans le cas singulier, nous montrons sous des hypothèses appropriées (Théorème 4.4.1) la convergence de $\mathbf{u}_\varepsilon^h$ vers $\mathbf{u}$ lorsque $\varepsilon \to 0$, $h/\varepsilon \to 0$ et $\varepsilon \ln h \to 0$. À cette fin, nous réutilisons lorsque cela est approprié les arguments du cas régulier, mais, pour gérer le point isolé dans le bord de Ω, nous devons aussi utiliser un équivalent en dimension deux ou trois de la *propriété d'engendrement* discutée dans la section 1.1.2, qui empêche la formation d'artéfacts en damier près de ce point.

Finalement, nous discutons l'approximation de la distance de Randers $\mathrm{dist}_{\mathcal{F}}(x, y)$ lorsque aucun des deux points x et y n'est fixé. Le schéma (1.12), modifié près de $\partial\Omega$ afin de prendre en compte la condition aux limites de Dirichlet $u_\varepsilon^h = 0$ sur $\partial\Omega$, peut être écrit sous la forme matricielle $L_\varepsilon^h u_\varepsilon^h = 0$, où L_ε^h est une matrice carrée indexée par x, $y \in \Omega \cap h\mathbb{Z}^d$. Nous montrons (Théorème 4.4.2) que, sous des hypothèses appropriées et localement uniformément par rapport à $(x, y) \in \Omega \times \Omega$,

$$-\varepsilon \ln[((L_\varepsilon^h)^{-1})_{xy}] \to \mathrm{dist}_{\mathcal{F}}(x, y), \qquad (\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0. \tag{1.16}$$

Lors de l'approximation de distances $\mathrm{dist}_{\mathcal{F}}(x, y)$ entre de nombreuses paires de points $(x, y) \in \Omega \times \Omega$, nous recommandons de préfactoriser la matrice L_ε^h de telle sorte que les éléments $((L_\varepsilon^h)^{-1})_{xy}$ puissent être calculés efficacement.

Application au transport optimal régularisé. Étant donné deux mesures de probabilité μ et ν supportées sur $\Omega \cap h\mathbb{Z}^d$, le problème de transport optimal de Wasserstein 1 entre μ et ν s'écrit

$$W(\mu, \nu) := \inf_{P \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \mathrm{dist}_{\mathcal{F}}(x, y)\, dP(x, y),$$

où $\Pi(\mu, \nu)$ désigne l'ensemble des *plans de transport* entre μ et ν, c'est-à-dire l'ensemble des mesures de probabilité sur $\Omega \times \Omega$ dont les première et seconde marginales coïncident respectivement avec μ et ν. Ce problème admet la régularisation entropique

$$W_\varepsilon(\mu, \nu) := \inf_{P \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \mathrm{dist}_{\mathcal{F}}(x, y)\, dP(x, y) - \varepsilon\, \mathrm{Ent}(P),$$

où $\mathrm{Ent}(P) := -\sum_{x,y\in\Omega\cap h\mathbb{Z}^d} P_{xy}\ln P_{xy}$ si $P = \sum_{x,y\in\Omega\cap h\mathbb{Z}^d} P_{xy}\delta_{(x,y)}$.

Il est bien connu que la régularisation entropique du problème de transport optimal peut être résolue en utilisant l'*algorithme de Sinkhorn* [Cut13]. Cet algorithme implique d'effectuer plusieurs produits matrice vecteur impliquant la matrice carrée K_ε indexée par x, $y \in \Omega \cap h\mathbb{Z}^d$ et définie par

$$(K_\varepsilon)_{xy} := \exp\left(-\frac{\mathrm{dist}_{\mathcal{F}}(x,y)}{\varepsilon}\right).$$

Puisque la matrice K_ε est dense, il serait extrêmement couteux numériquement de la calculer explicitement et il est donc souhaitable de plutôt trouver une approximation de K_ε pour laquelle les produits matrice vecteur peuvent être calculés efficacement. D'après (1.16), nous recommandons d'approcher K_ε par $(L_\varepsilon^h)^{-1}$. Les produits matrice vecteur impliquant $(L_\varepsilon^h)^{-1}$ se réduisent à des systèmes linéaires creux, qui peuvent être résolus efficacement en particulier si la matrice L_ε^h a été préfactorisée au préalable.

Une approche similaire a été introduite précédemment dans [Sol+15] dans le cadre des variétés riemanniennes. Nous remarquons que, alors que le problème de transport optimal dans lequel le coût de transport $\mathrm{dist}_{\mathcal{F}}(x,y)$ a été remplacé par son carré $\mathrm{dist}_{\mathcal{F}}(x,y)^2$ est aussi discuté dans [Sol+15], l'approche utilisée pour gérer cette variante quadratique ne peut pas être généralisée au cadre des variétés de Randers.

Nous affichons dans la Figure 1.6 quelques résultats numériques obtenus en résolvant des problèmes de transport optimal en utilisant la méthode proposée, dans des variétés de Randers associées au problème de navigation de Zermelo.

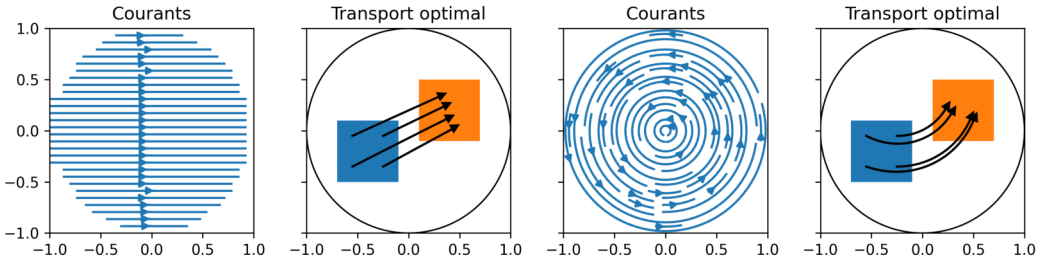


FIGURE 1.6 : Transport optimal dans le cadre du problème de navigation de Zermelo. À gauche et au milieu à droite : courants ambiants. Au milieu à gauche et à droite : à chaque point du domaine source (en bleu), le plan de transport optimal associe une mesure dans le domaine cible (en orange). Nous affichons des flèches pointant vers les barycentres de ces mesures. Le fait que le transport optimal ne soit pas une translation même lorsque les courants sont constants est une propriété connue du problème de Wasserstein 1.

1.2.2 Un schéma monotone et consistant à l'ordre deux pour l'équation de Pucci bidimensionnelle

Dans le Chapitre 5, nous introduisons une discrétisation monotone de l'équation de Pucci bidimensionnelle

$$\lambda_{\min}(D^2u(x)) + \mu\lambda_{\max}(D^2u(x)) = f(x), \tag{1.17}$$

où $\lambda_{\min}$ et $\lambda_{\max}$ désignent respectivement la plus petite et la plus grande valeur propre et où $\mu > 0$ est un paramètre donné. Pour tout $\theta \in \mathbb{R}$, nous définissons la matrice de rotation R_θ et la

matrice pivotée $\mathcal{D}(\theta, \mu)$ par

$$R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \mathcal{D}(\theta, \mu) := R_\theta \begin{pmatrix} 1 & 0 \\ 0 & \mu \end{pmatrix} R_\theta^\top;$$

alors l'équation de Pucci (1.17) admet la reformulation suivante sous la forme (1.2) :

$$\max_{\theta \in [0,\pi]} -\operatorname{Tr}(\mathcal{D}(\theta, \mu) D^2 u(x)) = -f(x), \tag{1.18}$$

pourvu que $\mu \leq 1$ (si $\mu \geq 1$, alors le maximum doit être remplacé par un minimum). En suivant l'approche décrite dans la section 1.1 pour discrétiser la reformulation (1.18), nous obtenons le schéma aux différences finies

$$\max_{\theta \in [0,\pi]} -\sum_{e \in \mathbb{Z}^d \setminus \{0\}} \lambda^e(\theta, \mu) \frac{u(x + he) + u(x - he) - 2u(x)}{h^2} = -f(x), \tag{1.19}$$

où, pour une valeur donnée de $\mu > 0$, les coefficients $\lambda^e(\theta, \mu) \geq 0$ sont non nuls uniquement pour un nombre fini de directions $e \in \mathbb{Z}^d \setminus \{0\}$ et peuvent être calculés en utilisant l'*algorithme de Selling*, voir la section 1.1.1.

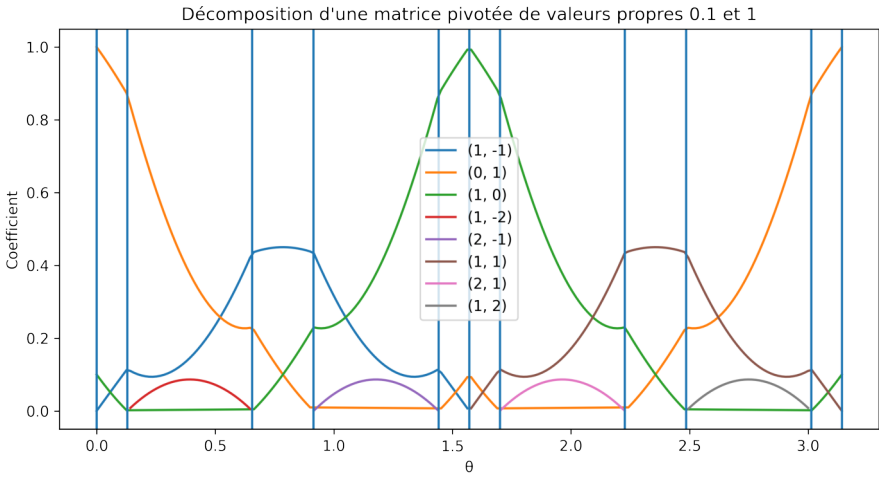


FIGURE 1.7 : Coefficients $\lambda^e(\theta, 0.1)$, pour différentes valeurs du vecteur direction e et de l'angle θ.

Une approche usuelle pour évaluer le membre de gauche dans le schéma (1.19) est de discrétiser l'espace des paramètres $[0, \pi]$ du maximum, ce qui introduit une erreur d'approximation. Ensuite, pour chaque θ dans l'ensemble des paramètres discrétisé, les coefficients $\lambda^e(\theta, \mu)$ non nuls et les directions e correspondantes peuvent être calculés en utilisant l'algorithme de Selling. Cette approche est coûteuse numériquement : si la grille cartésienne sur laquelle le schéma (1.19) est posé a un nombre d'éléments de l'ordre de N^2, alors le nombre d'éléments de la discrétisation de l'ensemble des paramètres doit aussi être de l'ordre de N^2 afin de préserver la consistance du schéma à l'ordre deux, auquel cas le coût numérique de l'évaluation du membre de gauche dans (1.19) en tous les points de la grille cartésienne est de l'ordre de N^4.

Notre principale contribution est de montrer (section 5.2.2) que le maximum dans (1.19) admet une expression de forme fermée. Le schéma peut donc être évalué sans discrétiser l'ensemble des paramètres, le coût numérique de cette évaluation sur la grille cartésienne introduite ci-dessus

étant seulement de l'ordre de N^2. Afin de dériver l'expression de forme fermée, nous montrons que, comme illustré par la Figure 1.7, les fonctions $\theta \mapsto \lambda^e(\theta, \mu)$ coïncident avec la somme d'une sinusoïde et d'une constante sur chacun d'un nombre fini d'intervalles fermés dont l'union est $[0, \pi]$. Chacun de ces intervalles $I \subset [0, \pi]$ est associé à un ensemble de matrices symétriques définies positives de trace normalisée

$$\left\{(1+\mu)^{-1}\mathcal{D}(\theta, \mu) \mid \theta \in I\right\}$$

qui est inclus dans une cellule unique de la triangulation infinie de la Figure 1.2.

Nous concluons le Chapitre 5 en effectuant quelques expériences numériques illustrant la précision de la méthode numérique que nous recommandons.

Nous appliquons à l'équation de Pucci l'idée de calculer une expression de forme fermée pour le maximum survenant dans un schéma aux différences finies conçu en utilisant les outils de la section 1.1, mais cette approche peut aussi être généralisée à d'autres équations aux dérivées partielles dégénérées elliptiques. Nous discutons le cas de l'équation de Monge-Ampère dans la section 1.2.3.

1.2.3 Discrétisation monotone de l'équation de Monge-Ampère du transport optimal

Dans le Chapitre 6, nous introduisons une discrétisation aux différences finies monotones du second problème aux limites pour l'équation de Monge-Ampère. C'est le problème aux limites pertinent dans le cas d'équations de Monge-Ampère associées à des problèmes de transport optimal. Nous considérons des équations de Monge-Ampère de la forme

$$\det\left(D^2u(x) - A(x, Du(x))\right) = B(x, Du(x)), \tag{1.20}$$

où $B \geq 0$ et où les matrices $A(x, Du(x))$ sont symétriques, notre résultat de convergence étant seulement prouvé dans le cas particulier d'équations de la forme

$$\det D^2u(x) = \frac{f(x)}{g(Du(x))}, \tag{1.21}$$

qui correspond à des problèmes de transport optimal dont la fonction coût est le carré de la distance Euclidienne (que nous appelons problèmes de transport optimal *quadratiques*). Les solutions de l'équation de Monge-Ampère sont considérées admissibles lorsqu'elles sont convexes, dans le cadre de l'équation (1.21), ou lorsqu'elles satisfont formellement la condition de convexité généralisée $D^2u(x) \succeq A(x, Du(x))$ pour l'ordre de Loewner sur les matrices symétriques, dans le cadre de l'équation (1.20).

Discrétisation. La forme générale du schéma numérique que nous recommandons étant décrite dans le Chapitre 6, considérons ici, pour des raisons de simplicité, la discrétisation de l'équation plus simple

$$\det D^2u(x) = f(x).$$

Cette équation admet la reformulation suivante, précédemment utilisée dans [FJ17] pour la conception d'un schéma numérique, sous la forme (1.2) :

$$\max_{\mathcal{D}\in\mathbf{S}_1}\left(df(x)^{1/d}(\det\mathcal{D})^{1/d} - \operatorname{Tr}(\mathcal{D}D^2u(x))\right) = 0, \tag{1.22}$$

où d est la dimension du domaine de l'équation et $\mathbf{S}_1$ désigne l'espace des matrices symétriques semi-définies positives de taille d dont la trace est égale à un. Le schéma proposé est une discrétisation aux différences finies de la reformulation ci-dessus :

$$F^h_{\mathrm{MA}} u(x) = 0, \qquad F^h_{\mathrm{MA}} u(x) := \max_{\mathcal{D} \in \mathbf{S}^h_1} \left(df(x)^{1/d} (\det \mathcal{D})^{1/d} - \Delta^{\mathcal{D}}_h u(x) \right), \tag{1.23}$$

où $\mathbf{S}^h_1$ est une approximation appropriée de $\mathbf{S}_1$ et où $\Delta^{\mathcal{D}}_h u(x)$ est une discrétisation aux différences finies de $\mathrm{Tr}(\mathcal{D} D^2 u(x))$ obtenue d'après (1.4), avec un choix approprié de coefficients $\lambda_i \geq 0$ et de directions $e_i \in \mathbb{Z}^d$ satisfaisant (1.3).

Un avantage d'utiliser la reformulation (1.22) de l'équation de Monge-Ampère est que l'équation (1.22) assure la convexité de ses solutions. En conséquence, de grands pas peuvent être utilisés lors de la résolution avec la méthode de Newton du schéma aux différences finies associé. À titre de comparaison, lors de la résolution de certains schémas aux différences finis précédemment introduits pour l'équation de Monge-Ampère, tels que [BCM16], des pas extrêmement petits doivent être utilisés dans la méthode de Newton afin de préserver la convexité des itérées.

En dimension $d = 2$, nous recommandons de choisir les coefficients λ_i et les vecteurs e_i dans $\Delta^{\mathcal{D}}_h u(x)$ comme ceux de la *décomposition de Selling* de la matrice $\mathcal{D}$, de façon cohérente avec l'approche décrite dans la section 1.1. Nous recommandons aussi de choisir l'ensemble $\mathbf{S}^h_1 \approx \mathbf{S}_1$ comme une sous-triangulation finie de la triangulation infinie du disque $\mathbf{S}_1$ affichée dans la Figure 1.2. Comme dans le cas de l'équation de Pucci (voir la section 1.2.2), nous prouvons (Théorème 6.1.2) que le maximum dans (1.23) admet alors une expression de forme fermée, la preuve utilisant le fait que les vecteurs e_i associés à des coefficients λ_i non nuls restent constants sur chaque cellule de la triangulation $\mathbf{S}^h_1$. Grâce à cette expression de forme fermée, l'opérateur discret F^h_{MA} dans (1.23) peut être évalué particulièrement efficacement.

	Cas général	Cas régulier, avec Lax-Friedrichs	Cas régulier, sans Lax-Friedrichs
Erreur de consistance	$O(h^{2/3})$	$O(h)$	$O(h^2)$
Coût numérique	$O(h^{-8/3} \log(1 + h^{-1}))$	$O(h^{-2})$	$O(h^{-2})$
Coût numérique (maximum discrétisé)	$O(h^{-10/3})$	$O(h^{-6})$	$O(h^{-4})$

TABLE 1.3 : Analyse de l'erreur de consistance et du coût numérique de la discrétisation aux différences finies proposée pour l'équation de Monge-Ampère.

Dans la Table 1.3, nous affichons l'ordre de consistance du schéma (1.23) avec l'équation de Monge-Ampère reformulée (1.22) et le coût numérique de l'évaluation de l'opérateur F^h_{MA}. À titre de comparaison, nous affichons aussi une estimation optimiste du coût numérique de l'évaluation de l'opérateur F^h_{MA} lorsqu'une discrétisation de l'ensemble des paramètres du maximum est utilisée plutôt que l'expression de forme fermée et lorsque cette discrétisation de l'ensemble des paramètres doit être suffisamment fine afin de préserver l'ordre de consistance du schéma. Notre analyse s'applique aussi à la généralisation du schéma (1.23) à des équations de Monge-Ampère de la forme générale (1.20). Nous distinguons entre trois cas différents. Dans le *cas général*, une partie de l'erreur de consistance est due à l'approximation de l'ensemble des paramètres $\mathbf{S}_1$ du maximum par $\mathbf{S}^h_1$ et le nombre de cellules dans la triangulation $\mathbf{S}^h_1$ doit augmenter lorsque le pas de discrétisation h décroît, ce qui a un effet sur le coût numérique de la méthode. Dans le *cas régulier*, nous supposons que la solution de l'équation de Monge-Ampère est régulière et satisfait une variante uniforme stricte de la condition de convexité généralisée $D^2 u(x) \succeq A(x, Du(x))$:

il est alors possible de choisir l'ensemble $\mathbf{S}_1^h$ indépendamment de h, ce qui améliore l'ordre de consistance du schéma tout en réduisant son coût numérique. Lors de la discrétisation d'équations de Monge-Ampère de la forme (1.20) pour lesquelles l'une au moins des fonctions A et B dépend effectivement de sa seconde variable, nous recommandons d'utiliser une approximation de Lax-Friedrichs du gradient $Du(x)$, consistante uniquement à l'ordre un ; c'est pourquoi nous distinguons deux sous-cas du cas régulier, selon la nécessité ou non d'utiliser une telle approximation de Lax-Friedrichs.

La condition aux limites de transport optimal. Nous supposons à partir de maintenant que l'équation de Monge-Ampère est posée sur un domaine borné $X \subset \mathbb{R}^d$. Dans le cadre (1.21) des équations de Monge-Ampère associées à des problèmes de transport optimal quadratiques, le second problème aux limites pour l'équation de Monge-Ampère implique la condition aux limites de transport optimal

$$\overline{Du(X)} = \overline{Y}, \tag{1.24}$$

où Y est un ensemble convexe donné. L'égalité ci-dessus est appelée une condition aux limites bien qu'elle implique tout le domaine X : c'est parce que sous des conditions appropriées, par un argument de convexité, elle peut être reformulée comme $Du(\partial X) = \partial Y$.

Le second problème aux limites pour l'équation de Monge-Ampère présente une propriété d'*invariance additive* : à la fois l'équation de Monge-Ampère et la condition aux limites de transport optimal impliquent les dérivées de l'inconnue u, mais pas ses valeurs elles-mêmes. En conséquence, l'ensemble des solutions du problème aux limites est stable par addition d'une constante.

La propriété d'invariance additive est une source de difficulté lors de la conception de schémas numériques. En particulier, elle peut être un obstacle à l'existence de solutions à ces schémas. Illustrons ce phénomène en prenant comme exemple un autre système d'équations additivement invariant :

$$\begin{cases} -u''(x) + f(x) = 0 & \text{dans } (-1,1), \\ u'(-1) = u'(1) = 0. \end{cases} \tag{1.25}$$

Il s'agit de l'équation de Poisson sur le domaine unidimensionnel $(-1,1)$, équipée d'une condition aux limites de Neumann. Ce système admet des solutions uniquement lorsque la condition de compatibilité suivante entre le terme source f et la condition aux limites est satisfaite :

$$\int_{-1}^{1} f(x)\,dx = 0. \tag{1.26}$$

Une condition similaire doit être satisfaite pour que le système de Monge-Ampère admette des solutions. Il s'agit de la condition d'équilibre des masses

$$\int_X f(x)\,dx = \int_Y g(y)\,dy. \tag{1.27}$$

La difficulté, dans le cadre des schémas aux différences finies, est que, souvent, aucun équivalent discret de (1.26) ou de (1.27) n'est satisfait. Une discrétisation aux différences finies naturelle du système (1.25) sur la grille $(-1,1) \cap h\mathbb{Z}^d$ serait

$$-\frac{\delta_h^+ u(x) + \delta_h^- u(x)}{h} + f(x) = 0 \quad \text{dans } (-1,1) \cap h\mathbb{Z},$$
$$\delta_h^\pm u(x) := \begin{cases} \frac{u(x\pm h) - u(x)}{h} & \text{si } u(x \pm h) \in (-1,1), \\ 0 & \text{sinon.} \end{cases}$$

Ce schéma n'admet pas de solutions, sauf dans le cas particulier d'un terme source f satisfaisant la condition de compatibilité discrète

$$\sum_{x\in(-1,1)\cap h\mathbb{Z}} f(x) = 0.$$

Le schéma ci-dessus peut être adapté afin d'admettre des solutions, mais toutes les adaptations envisageables ne peuvent pas être généralisées à des discrétisations aux différences finies de l'équation de Monge-Ampère.

La discrétisation que nous introduisons pour la condition aux limites de transport optimal (1.24) est fondée sur le fait que l'inclusion $Du(X) \subset \overline{Y}$ puisse être réécrite comme l'inégalité

$$\max_{|e|=1} (\langle e, Du(x)\rangle - \sigma_Y(e)) \leq 0 \quad \text{dans } X,$$

où $\sigma_Y : e \mapsto \sup_{p\in P(x)}\langle e, p\rangle$ est la *fonction support convexe* de l'ensemble $P(x)$. Cette reformulation peut être discrétisée en utilisant des différences finies décentrées amont standard. Il en résulte un schéma de la forme

$$F^h_{\mathrm{BV2}}u(x) \leq 0 \qquad\qquad F^h_{\mathrm{BV2}}u(x) := \max_{|e|=1} (D^e_h u(x) - \sigma_{P(x)}(e)). \tag{1.28}$$

Nous avons à présent une égalité (1.23) et une inégalité (1.28), toutes deux posées sur une discrétisation cartésienne du domaine X. Notre objectif est de concevoir un schéma numérique pour le système d'équation complet du second problème aux limites, en utilisant chacun des opérateurs discrets F^h_{MA} et F^h_{BV2}.

Une façon d'y parvenir est d'adapter à notre cadre l'approche développée dans [Fro19]. Il en résulterait le schéma

$$\max\{F^h_{\mathrm{MA}}u(x), F^h_{\mathrm{BV2}}u(x)\} = 0, \tag{1.29}$$

modifié près de ∂X pour prendre en compte la condition aux limites de Dirichlet $u = 0$ sur ∂X. Cette approche a été étudiée dans [Fro19] dans le cadre (1.21) des problèmes de transport optimal quadratiques, bien qu'avec des définitions différentes des opérateurs F^h_{MA} et F^h_{BV2}. Le fait de prendre le maximum entre les opérateurs F^h_{MA} et F^h_{BV2} y est justifié par un phénomène de compétition entre les inégalités $F^h_{\mathrm{MA}}u(x) \leq 0$ et $F^h_{\mathrm{BV2}}u(x) \leq 0$. La condition aux limites de Dirichlet est satisfaite dans au sens classique près d'un point $x_* \in \partial X$, alors qu'aux autres points elle doit être comprise dans le sens faible $u \leq 0$. Son objectif est double : d'une part, l'égalité $u(x_*) = 0$ sélectionne une unique solution parmi l'ensemble additivement invariant des solutions du problème de Monge-Ampère et, d'autre part, elle affaiblit la discrétisation de la condition aux limites de transport optimal près du point x_*, affaiblissant donc aussi la nécessité d'un équivalent discret de la condition d'équilibre des masses (1.27). Un ingrédient clé dans l'analyse menée dans [Fro19] est que le schéma numérique satisfait une propriété de *sous-estimation*. Une difficulté survenant lorsque l'on tente d'étendre cette analyse à des équations de Monge-Ampère de la forme générale (1.20) est le manque de stratégie évidente permettant de discrétiser (1.20) de façon *sous-estimatrice*.

Afin d'éviter d'avoir à utiliser la propriété de sous-estimation et aussi afin d'éviter certains artéfacts numériques qui tendent à se produire près du point x_*, nous utilisons une légère modification du schéma (1.29). De façon similaire aux expériences numériques dans [BD19], nous ajoutons une inconnue $\alpha \in \mathbb{R}$ au problème discret et nous résolvons

$$\max\{F^h_{\mathrm{MA}}u(x) + \alpha, F^h_{\mathrm{BV2}}u(x)\} = 0. \tag{1.30}$$

La présence de l'inconnue α suffit à affaiblir la nécessité d'un équivalent discret à la condition d'équilibre des masses (1.27). Nous remplaçons la condition aux limites $u = 0$ par $u = +\infty$ sur

∂X. Le schéma (1.30) assure seulement cette condition aux limites de Dirichlet dans le sens faible $u \leq +\infty$ — ce qui est toujours vérifié — sur l'ensemble du bord ∂X. Nous montrons que cette condition aux limites de Dirichlet n'entraîne pas de couche limite. Afin de sélectionner une unique solution parmi l'ensemble additivement invariant des solutions au problème de Monge-Ampère, nous ajoutons une contrainte $u(x_0) = 0$, pour un point $x_0 \in X$ donné. Ajouter une contrainte d'égalité est cohérent avec le fait d'ajouter une inconnue α au schéma.

Un désavantage à ajouter l'inconnue α est que, bien que les opérateurs F^h_{MA} et F^h_{BV2} soient tous les deux monotones par rapport à leur argument u, le schéma (1.30) n'est pas monotone par rapport à la paire d'inconnues (u, α). La méthode de Perron, qui permet souvent de prouver l'existence de solutions à des schémas numériques monotones, ne s'applique donc pas directement au schéma (1.30). Nous montrons (Théorème 6.2.14) que la méthode de Perron peut cependant être adaptée à notre cadre, en gérant l'inconnue α séparément de u dans la preuve de l'existence de solutions. Nous énonçons notre preuve comme un résultat général à propos de l'existence de solutions à une classe de schémas numériques additivement invariant dont les inconnues sont une fonction discrète u et un scalaire $\alpha \in \mathbb{R}$ et qui sont monotones par rapport à u pour des valeurs fixées de α.

Sous des hypothèses appropriées, nous établissons la convergence du schéma (1.30) dans le cadre (1.21) des problèmes de transport optimal quadratiques (Théorème 6.5.22). Dans ce but, nous devons étudier le lien entre, d'une part, les sous-solutions et sur-solutions de *viscosité* du système d'équations discrétisé et, d'autre part, les solutions d'*Aleksandrov* du problème de Monge-Ampère. Le cas des sous-solutions de viscosité a été étudié précédemment dans le cadre de [Fro19], mais pas le cas des sur-solutions de viscosité, dont l'analyse n'était pas nécessaire dans ce cadre.

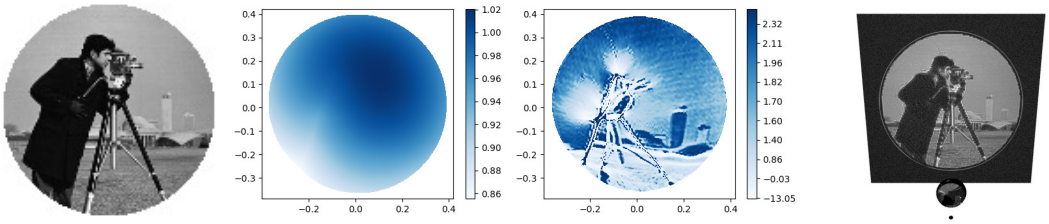


FIGURE 1.8 : Application de la méthode au problème du réfracteur en champ lointain en optique non imageante. À gauche : image cible. Au milieu à gauche : forme de la lentille, calculée en utilisant une implémentation de la méthode avec le langage de programmation Python®. Au milieu à droite : approximation de la courbure ponctuelle de la lentille. À droite : simulation de la scène, en utilisant le moteur de rendu appleseed®. Le petit point noir en bas représente la source de lumière.

Bien que la convergence du schéma (1.30) reste un problème ouvert dans le cas d'équations de Monge-Ampère de la forme générale (1.20), nous avons effectué des expériences numériques afin de valider la méthode dans le cadre du problème du réfracteur en champ lointain en optique non imageante. Dans ce problème, des rayons de lumière émanent d'une source de lumière ponctuelle, ils sont déviés par réfraction par une lentille optique et les rayons déviés continuent de se propager jusqu'à ce qu'ils rencontrent un écran, dont la distance à la source ponctuelle et à la lentille est supposée grande. L'objectif est de trouver une forme appropriée pour la lentille afin qu'une image cible donnée soit projetée sur l'écran. Il est bien connu [GH09] que ce problème est décrit par une équation de Monge-Ampère de la forme (1.20). Nous résolvons le schéma numérique associé (1.30),

nous en déduisons une approximation de la forme de la lentille et nous simulons la propagation de la lumière pour la forme que nous obtenons, comme illustré par la Figure 1.8. Nous observons que le graphe de la courbure ponctuelle approchée de la lentille rappelle l'image cible, mais que les zones claires de l'image cible tendent à être agrandies alors que les zones sombres tendent à être rétrécies, ce qui est un résultat attendu puisque les parties de la lentille correspondant aux zones claires doivent être grandes afin de capturer plus de rayons de lumière.

Chapter 1

Introduction

In this thesis, we study the monotone finite difference discretization on Cartesian grids of *degenerate elliptic* partial differential equations, that is, equations of the form

$$F(x, u(x), Du(x), D^2u(x)) = 0,$$

where the function F is nondecreasing with respect to its second variable and nonincreasing with respect to its last variable. The main focus is on the discretization of the anisotropic non-divergence form diffusion operator

$$u \mapsto \mathrm{Tr}(\mathcal{D}(\cdot)D^2u(\cdot)), \tag{1.1}$$

where $\mathcal{D}(\cdot)$ is a field of symmetric positive definite matrices. Degenerate elliptic equations featuring such a diffusion operator may for instance take the form

$$\sup_{\alpha\in\mathcal{A}} \left(H^\alpha(x, u(x), Du(x)) - \mathrm{Tr}(\mathcal{D}^\alpha(x)D^2u(x))\right) = 0, \tag{1.2}$$

where $\mathcal{A}$ is a given parameter set, functions H^α are nondecreasing with respect to their second variables, and $\mathcal{D}^\alpha$ are fields of symmetric positive definite matrices. The Pucci and Monge-Ampère equations admit reformulations in the form (1.2), see sections 1.2.2 and 1.2.3.

The property of *degenerate ellipticity* admits a discrete counterpart, that we call *monotonicity* or *discrete degenerate ellipticity* depending on the exact variant of this counterpart that we consider. In many situations, a discrete comparison principle may be proved in a very straightforward way for solutions to discrete degenerate elliptic numerical schemes. Moreover, a general argument was introduced in [BS91] for proving convergence of monotone numerical schemes for degenerate elliptic partial differential equations. For all those reasons, it is desirable to design discretizations of degenerate elliptic equations that satisfy the property of monotonicity.

The finite difference discretization of the diffusion operator (1.1) that is considered in this thesis is based on a decomposition of any symmetric positive definite matrix $\mathcal{D}$ in the form

$$\mathcal{D} = \sum_{i=1}^{I} \lambda_i e_i e_i^\top, \tag{1.3}$$

where coefficients λ_i are nonnegative and offsets $e_i \in \mathbb{Z}^d$ have integer elements. This yields the second-order consistent approximation

$$\mathrm{Tr}(\mathcal{D}D^2u(x)) \approx \sum_{i=1}^{I} \lambda_i \frac{u(x + he_i) + u(x - he_i) - 2u(x)}{h^2}. \tag{1.4}$$

The elements of the offsets e_i must be integers so that only points of a Cartesian grid are involved in the finite difference discretization. This constraint prevents simply using the eigendecomposition of the matrix $\mathcal{D}$. The coefficients λ_i must be nonnegative in order for the discretization to be suitable for use in numerical schemes that have to satisfy the monotonicity property.

The above strategy is not the only way to build a monotone and consistent discretization of the operator (1.1). We could for instance have chosen to relax the constraint that offsets e_i have integer elements, using instead an interpolation step in the numerical scheme in order to approximate values of the unknown at points that do not belong to the discretization grid. This approach is well-known [DJ13; NNZ19] and yields a two-scale discretization of the original operator. We choose to retain the constraint that e_i have integer elements, with the hope of obtaining finite difference stencils with smaller radii and schemes with higher orders of consistency.

Historically, the numerical analysis of schemes based on a decomposition in the form (1.3) satisfying the above constraints was performed [Kry05; KT92] even though no method to compute this decomposition in a numerically practical way was known, except for some specific classes of matrices $\mathcal{D}$ such as diagonally dominant matrices. A construction based on the Stern-Brocot tree of irreducible fractions was introduced in [BOZ04] for matrices of size two, and another construction, equivalent for matrices of size two but also applicable for matrices of size three, and based on Selling's algorithm [CS92; Sel74], a tool originating from the study of low-dimensional lattice geometry, was independently introduced in [FM14]. The second construction has a natural extension to symmetric positive definite matrices of size greater than three, based on Voronoi's first reduction of quadratic forms, which is well-known in the theory of low-dimensional lattice geometry [CS88; Vor08] and was previously used for the discretization of partial differential equations in [Mir19].

The first part of this thesis is devoted to the study of some properties of Voronoi's first reduction that are of interest when applying it to the design of finite difference discretizations. The second part is devoted to the study of finite difference schemes for some specific degenerate elliptic equations of the form (1.2).

1.1 Tools for the monotone finite difference discretization of anisotropic differential operators on Cartesian grids

If $\mathcal{D}$ is a given symmetric positive definite matrix, we recommend choosing the coefficients of its decomposition (1.3) as a solution to the maximization problem

$$\max \left\{ \sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e \;\middle|\; \lambda\colon \mathbb{Z}^d\setminus\{0\} \to \mathbb{R}_+, \sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e ee^\top = \mathcal{D} \right\}. \tag{1.5}$$

While this problem is posed on the set of all, possibly infinitely supported, decompositions of $\mathcal{D}$ with nonnegative coefficients and offsets with integer elements, it may be shown that its solutions are always finitely supported. Moreover, since the constraint $\sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e ee^\top = \mathcal{D}$ implies that $\sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e |e|^2 = \mathrm{Tr}(\mathcal{D})$, maximizing the coefficients λ^e is consistent with the fact that it is desirable, for numerical purposes, that offsets $e \in \mathbb{Z}^d \setminus \{0\}$ associated with nonzero coefficients λ^e have small norms.

The maximization problem (1.5) admits the following dual problem:

$$\min_{M\in\mathcal{M}_d} \mathrm{Tr}(\mathcal{D}M); \quad \mathcal{M}_d := \{M \in \mathbb{R}^{d\times d} \text{ symmetric}, \langle e, Me\rangle \geq 1, \forall e \in \mathbb{Z}^d \setminus \{0\}\}. \tag{1.6}$$

The admissible set $\mathcal{M}_d$ of this dual problem happens to be well-known in the field of low-dimensional lattice geometry, under the name of *Ryskov's polyhedron*. Although it is defined by infinitely many constraints, it features locally the same structure as a standard polyhedron [Sch09a].

In order to explain the relation between Ryskov's polyhedron and low-dimensional lattice geometry, let us associate to any symmetric positive definite matrix M of size d the lattice, unique up to a rotation, spanned by linear combinations with integer coefficients of the columns of some matrix B of size d such that $B^\top B = M$. Then Ryskov's polyhedron is the set of all symmetric positive definite matrices associated to lattices for which a nonoverlapping sphere packing may be built by placing a sphere of radius one half centered at each point of the lattice, as illustrated by Figure 1.1. Vertices of Ryskov's polyhedron are called *perfects forms*, and associated lattices are called *perfect lattices*.

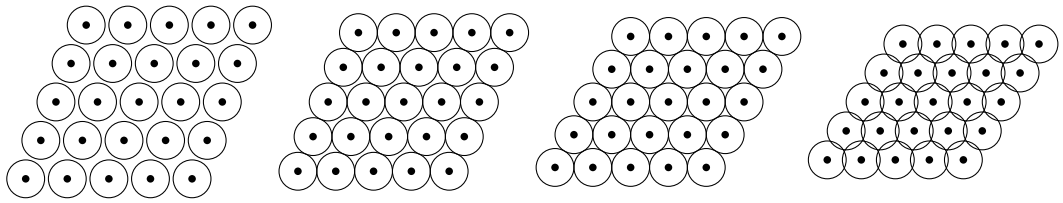


Figure 1.1: Ryskov's polyhedron: lattice interpretation. In the limit case of a perfect lattice, each sphere is in contact with at least $d(d+1)$ neighbor spheres (that is, 6 neighbor spheres in dimension $d = 2$).

While in applications to sphere packings the quantity that is minimized over Ryskov's polyhedron usually is the determinant, rather than some linear form as in the problem (1.6), both minimization problems share the property that their minimum is attained at some perfect form, that is, at some vertex of Ryskov's polyhedron. This motivated the classification of perfect forms that has been performed in the literature up to dimension $d = 8$ [CS88; DSV07]. The classification of symmetric positive definite matrices $\mathcal{D}$ depending on which perfect form M is optimal in (1.6) is called *Voronoi's first reduction of quadratic forms*. Two perfect forms M and M' are called *arithmetically equivalent* if they are associated to the same lattice, or equivalently if there exists some matrix A of size d with integer elements and whose inverse also has integer elements such that $M' = A^\top M A$. While, in any dimension $d \geq 2$, there are infinitely many perfect forms, Voronoi proved that there are only finitely many equivalence classes of perfect forms for the relation of arithmetical equivalence [Vor08]. Knowing those equivalence classes, as well as the way that the relation of arithmetical equivalence interacts with the polyhedral structure of Ryskov's polyhedron, allows one to solve the dual problem (1.6), and thus also the primal problem (1.5), in a particularly efficient manner, by walking over the adjacency graph of perfect forms.

Note that the number of equivalence classes of perfect forms explodes as the dimension increases, as illustrated by Table 1.1. Therefore the approach described in this thesis for building decompositions in the form (1.3) is tractable provided that the considered dimensions remains reasonably small.

In dimensions $d \in \{2, 3\}$, the structure of this classification is particularly simple since there exists only one equivalence class of perfect forms for the relation of arithmetical equivalence. Thus the minimization problem (1.6) may be solved using tools that are specific to those dimensions,

$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$	$d\geq 9$
1	1	2	3	7	33	10916	unknown

Table 1.1: Number of equivalence classes of perfect forms depending on the dimension.

such as Selling's algorithm [CS92; Sel74], which has been used previously for the discretization of partial differential equations [FM14]. In Chapter 2, we show how viewing the decomposition (1.3) obtained using Selling's algorithm as a solution to the maximization problem (1.5) allows to prove some theoretical guarantees regarding the feasibility of the second-order consistent monotone discretization of a differential operator involving both a first- and a second-order terms.

In Chapter 3, we explain how one may efficiently compute the set of solutions to the maximization problem (1.5) in dimension $d=4$, we recommend a way to choose a particular solution when this set is not a singleton, and we discuss some properties of the resulting finite difference schemes, namely the Lipschitz regularity of their coefficients, the radius of their stencils and the absence of checkerboard artifacts.

1.1.1 Second order monotone finite differences discretization of linear anisotropic differential operators

In dimensions $d \in \{2,3\}$, all perfect forms are equivalent, for the relation of arithmetical equivalence discussed above, to the reference perfect form

$$\frac{1}{2}\begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 2 \end{pmatrix} = \frac{1}{2}I_d + \frac{1}{2}\mathbb{1}\mathbb{1}^\top. \tag{1.7}$$

For this reason, all decomposition in the form (1.3), obtained using Voronoi's first reduction of quadratic forms, of all symmetric positive definite matrix $\mathcal{D}$ of size two of three have sets of offsets featuring the same particular structure, related to the notion of superbase of the lattice $\mathbb{Z}^d$ (a basis of $\mathbb{Z}^d$ is a family $v=(v_1,\dots,v_d)$ of vectors with integer elements satisfying $\det(v_1,\dots,v_d)=\pm 1$; a superbase of $\mathbb{Z}^d$ is a basis of $\mathbb{Z}^d$ extended with the additional vector $v_0 := -v_1-\dots-v_d$). More precisely, there exists a superbase v of $\mathbb{Z}^d$ such that the decomposition is supported by the offsets $\pm e_{ij}$, $i,\ j \in \{0,\dots,d\}$, defined by $\pm\langle e_{ij}, v_k\rangle = \pm(\delta_{ik}-\delta_{jk})$, for all $k \in \{0,\dots,d\}$. Moreover, the coefficient $\lambda^{e_{ij}}$ associated to the offset e_{ij} is equal to the scalar product $-\langle v_i, \mathcal{D}v_j\rangle$, and the constraint that it should be nonnegative may be seen as a property of *obtuseness* of the superbase v for the relevant scalar product.

We display in Figure 1.2 the sets of offsets in the decompositions of some symmetric positive definite matrices of size two. In dimension $d=2$, the set of symmetric positive definite matrices with trace normalized to one is a disk, which may be parametrized as

$$\left\{ \frac{1}{2}\begin{pmatrix} 1+\rho_1 & \rho_2 \\ \rho_2 & 1-\rho_1 \end{pmatrix} \;\middle|\; \rho_1^2+\rho_2^2 \leq 1 \right\}.$$

The offsets are constant on some triangles, which form an infinite triangulation of the disk and which coincide with cells in Voronoi's first reduction of the set of two-dimensional quadratic forms with normalized trace. In dimension two, the offsets of the decomposition themselves coincide, up to a change of sign, with the elements of some superbase of $\mathbb{Z}^2$.

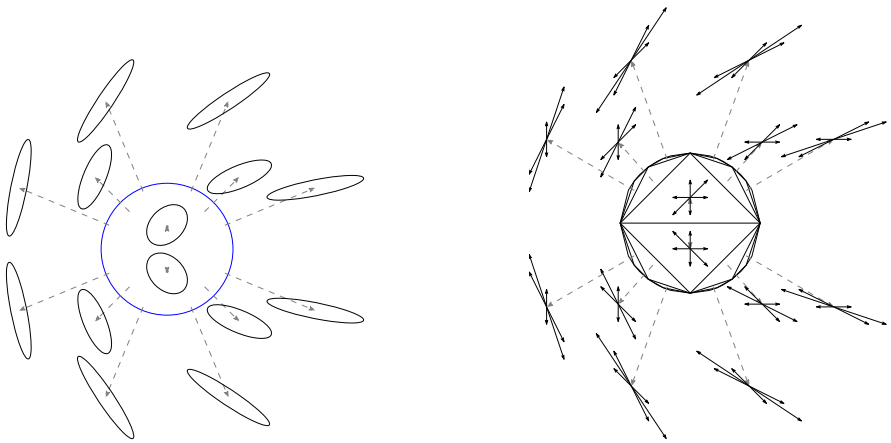


Figure 1.2: Left: the set of symmetric positive definite matrices of size two with unit trace. Right: the associated finite difference stencils.

While the finite difference discretization of second-order anisotropic differential operators in dimensions $d \in \{2, 3\}$, based on a decomposition of the coefficient matrix obtained using the particular structure of perfect forms in dimensions $d \in \{2, 3\}$, has been studied previously [FM14], we consider in Chapter 2 differential operators featuring both a first- and a second-order terms. At least two standard discretization strategies exist for the first-order term: the one using upwind finite differences, which are first-order consistent and monotone, and the one using centered finite differences, which are second-order consistent but lack monotonicity. In dimension one, it is well-known that, in presence of a nondegenerate second-order term in the discretized differential operator, the lack of monotonicity of the centered finite difference discretization of the first-order term may be compensated by the monotonicity of the discretization of the second-order term, allowing the whole operator to be discretized in a second-order consistent manner. We study how this construction may be extended to two- or three-dimensional differential operators of the form

$$u \mapsto \langle b(\cdot), Du(\cdot)\rangle + \mathrm{Tr}(\mathcal{D}(\cdot)D^2u(\cdot)),$$

where b and $\mathcal{D}$ are fields of respectively vectors and symmetric positive definite matrices.

We introduce the second-order consistent discretization

$$\begin{aligned}\langle b, Du(x)\rangle + \mathrm{Tr}(\mathcal{D}D^2u(x)) \approx &\sum_{i=1}^{I} \mu_i \frac{u(x+he_i) - u(x-he_i)}{2h} \\ &+ \sum_{i=1}^{I} \lambda_i \frac{u(x+he_i) + u(x-he_i) - 2u(x)}{h^2},\end{aligned} \tag{1.8}$$

where λ_i and e_i are respectively coefficients and offsets of a decomposition of the symmetric positive definite matrix $\mathcal{D}$ in the form (1.3), and μ_i are coefficients of a decomposition of the vector b in the form

$$b = \sum_{i=1}^{I} \mu_i e_i. \tag{1.9}$$

It is important that the same offsets e_i are shared between the decompositions of $\mathcal{D}$ and b, since otherwise the lack of monotonicity of the discretization of the first-order term could not be compensated.

We recommend choosing the decomposition (1.3) based on Voronoi's first reduction of quadratic forms, after having computed it using Selling's algorithm, and then choosing the coefficients μ_i according to the formula

$$\mu_i := \lambda_i \langle b, \mathcal{D}^{-1} e_i \rangle.$$

We show (Theorem 2.1.7) that with this choice of coefficients and offsets, the discretization (1.8) is monotone provided that the discretization step h is small enough, more precisely smaller than $C|\mathcal{D}^{-1}|^{-1/2}\langle b, \mathcal{D}^{-1} b\rangle^{-1/2}$, where the constant C depends only on the dimension. The main result in Chapter 2 is that this choice is quasi-optimal, in the sense that if the discretization (1.8) is monotone for some choice of coefficients μ_i, λ_i and of offsets e_i, then it is also monotone for the proposed choice of coefficients and offsets, up to dividing the discretization step h by a factor two in dimension two, or by a factor six in dimension three (Theorem 2.1.6).

In order to prove the main result, we have to study an extension of Voronoi's first reduction of quadratic forms to inhomogeneous quadratic forms. This extension involves the following variant of Ryskov's polyhedron:

$$\widetilde{\mathcal{M}}_d := \{(\eta, M) \mid \eta \in \mathbb{R}^d,\ M \in \mathbb{R}^{d\times d} \text{ symmetric};\ \forall e \in \mathbb{Z}^d \setminus \{0\},\ \langle \eta, e\rangle + \langle e, Me\rangle \geq 1\}.$$

While $\mathcal{M}_d$, the standard variant of Ryskov's polyhedron, is a set of symmetric matrices, the polyhedron $\widetilde{\mathcal{M}}_d$ that we introduce is a set of pairs of vectors and symmetric matrices. We show that the polyhedral structure of $\widetilde{\mathcal{M}}_d$ is remarkably similar to the one of Ryskov's polyhedron $\mathcal{M}_d$: for instance, all vertices of $\widetilde{\mathcal{M}}_d$ are of the form $(0, M)$, where M is a vertex of $\mathcal{M}_d$.

1.1.2 Monotone discretization of anisotropic four-dimensional differential operators using Voronoi's first reduction

In dimension $d = 4$, unlike in dimensions two and three, perfect forms are not necessarily arithmetically equivalent to the reference perfect form (1.7), and there exists exactly one other equivalence class of perfect forms, of which a representative is

$$\frac{1}{2}\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}. \tag{1.10}$$

Perfect forms that are arithmetically equivalent to (1.10) are not related to *superbases* of $\mathbb{Z}^d$ in the manner that those that are arithmetically equivalent to (1.7) are. Therefore, Selling's algorithm, which iterates over superbases of $\mathbb{Z}^d$, is not directly applicable in dimension $d = 4$. However, the minimization problem (1.6) may still be solved by walking over the vertices of Ryskov's polyhedron in a suitable manner. We explain in Chapter 3 that this procedure may be implemented particularly efficiently, using the fact that the adjacency relations between perfect forms, in the graph of vertices of Ryskov's polyhedron, may be precomputed.

We describe how one can deduce the set of solutions to the maximization problem (1.5). This set is not always a singleton. This is related to the fact that, as illustrated by Table 1.2, some perfect forms—those that are arithmetically equivalent to (1.10)—are degenerate vertices of Ryskov's polyhedron, meaning that they have more than ten neighbors, ten being the dimension of the space of symmetric matrices of size four. We show however that, in dimension $d = 4$ and up to identifying offsets $e \in \mathbb{Z}^4 \setminus \{0\}$ with their opposites, the set of solutions to (1.5) is always an equilateral triangle, possibly reduced to a singleton. We recommend choosing the barycenter of this set as the decomposition of the symmetric positive definite matrix $\mathcal{D}$ to be used in finite difference discretizations, and we prove some properties of the resulting discretizations:

d	perfect form arithmetically equivalent to	number of neighbors arithmetically equivalent to		dimension of the space of symmetric matrices of size d	degenerate perfect form
		(1.7)	(1.10)		
2	(1.7)	3	0	3	no
3	(1.7)	6	0	6	no
4	(1.7)	0	10	10	no
4	(1.10)	48	16	10	yes

Table 1.2: Polyhedral structure of Ryskov's polyhedron in dimensions $d \in \{2, 3, 4\}$.

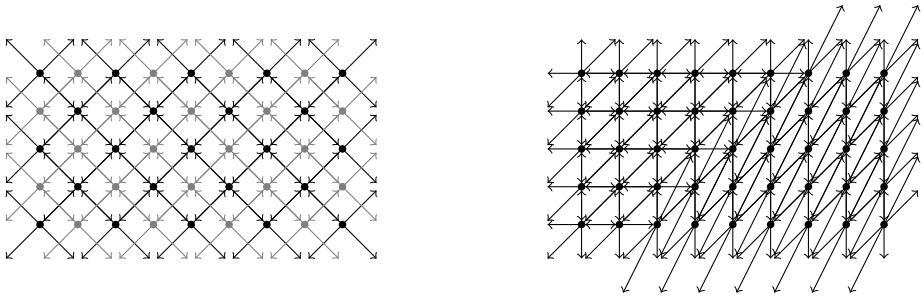
Lipschitz regularity of the coefficients. We show (Theorem 3.3.6) that the coefficients λ^e of the proposed decomposition of the symmetric positive definite matrix $\mathcal{D}$ depend on $\mathcal{D}$ in a locally Lipschitz manner, the Lipschitz constant depending only on the condition number of $\mathcal{D}$. As a corollary, when discretizing an anisotropic diffusion operator such as (1.1) involving a field $\mathcal{D}(\cdot)$ of symmetric positive definite matrices that is Lipschitz continuous and has bounded condition number, the fields of coefficients $\lambda^e(\cdot)$ in the discretization remain Lipschitz continuous. One of the benefits of proving continuity of the coefficients of the discretization is that it is an assumption in some well-known results about convergence rates of some numerical schemes for some degenerate elliptic partial differential equations [BJ07].

Radius of the finite difference stencil. The stencil of the finite difference discretization of an anisotropic diffusion operator associated to the decomposition of a symmetric positive definite matrix $\mathcal{D}$ that is solution to the maximization problem (1.5) is the set of offsets $e \in \mathbb{Z}^d \setminus \{0\}$ associated to a nonzero coefficient λ^e. For efficiency of the finite difference scheme, it is desirable that such offsets have small norm. We prove the estimate $|e| \leq C\mu(\mathcal{D})$, where C is a constant depending only on the dimension and $\mu(\mathcal{D}) := |\mathcal{D}|^{1/2}|\mathcal{D}^{-1}|^{1/2}$ is the square root of the condition number of the matrix $\mathcal{D}$ (Theorem 3.4.1). The value of the constant C is important and is discussed in section 3.4.1.

The above estimate is true in any dimension $d \in \mathbb{N}^*$ and for decompositions defined by any λ that is maximal in (1.5). It had previously only been proved in dimensions $d \in \{2, 3\}$, while in higher dimensions only the weaker estimate $|e| \leq C\mu(\mathcal{D})^{d-1}$ was known, see [Mir19]. In order to prove the improved estimate, we need to answer the following question: knowing that $\mathcal{D}$ belongs to some particular cell of Voronoi's first reduction of quadratic forms, is there a finite number of cells of this reduction to the union of which the matrix $\mathcal{D}^{-1}$ is guaranteed belong? We show that this is true, using that cells of Voronoi's first reduction are convex conical hulls of sets of rank one matrices, which greatly simplifies the structure of their images by the matrix inverse function.

Guarantees against checkerboard artifacts. Checkerboard artifacts typically occur when the graph of adjacency of points of a discretization grid, according to the stencils of a finite difference scheme, is not connected, as illustrated in Figure 1.3. Then the restrictions of the scheme to the different connected components of the grid behave independently, which may be undesirable. We present a strategy for proving the absence of such artifacts in some discretizations of four-dimensional operators involving anisotropic diffusion, using the previously proved properties on the Lipschitz continuity of the coefficients λ^e and the norm of the offsets e of the recommended decomposition of a symmetric positive definite matrix $\mathcal{D}$, as well as the following *spanning property*: the set of offsets e associated to nonzero coefficients λ^e spans the lattice $\mathbb{Z}^4$ by linear combinations with integer coefficients. We show (Theorem 3.5.1) that while there exist decompositions of $\mathcal{D}$ associated to maximizers λ of the problem (1.5) that do not satisfy this last property, the

recommended decomposition, which is associated to the barycenter of the set of maximizers, does satisfy it.



scheme with checkerboard artifacts scheme without checkerboard artifacts

Figure 1.3: Left: stencils in a discretization of the Laplace equation using diagonal offsets, which is an example of scheme featuring checkerboard artifacts. Right: stencils associated to the discretization recommended in this thesis for an anisotropic diffusion operator, featuring good connectivity properties which prevent checkerboard artifacts.

1.2 Monotone discretization of some specific degenerate elliptic partial differential equations

In the second part of this thesis, we study some specific finite difference schemes, designed using the previously described strategy for discretizing anisotropic diffusion operators. Those schemes allow us, on the one hand, to approximate Randers distances and associated optimal transport distances, and on the other hand, to solve numerically the Pucci and Monge-Ampère equations, reformulated in the form (1.2).

1.2.1 A linear finite-difference scheme for approximating Randers distances on Cartesian grids

Randers distances. Randers distances are an asymmetric extension to Riemannian distances. Their asymmetry may illustrate, for instance, the fact that it is easier to move downwind than upwind in a medium subject to currents, or easier to move downhill than uphill due to the effect of the gravity. Zermelo's navigation problem [BRS04], illustrated in Figure 1.4, is an example of a standard problem involving Randers distances.

Randers distances have been initially introduced in the setting of general relativity [Ran41]. They have numerous other applications, including image segmentation [CMC16], quantum vortices [ABM06], and path curvature penalization [CMC17].

A Randers metric in an open domain $\Omega \subset \mathbb{R}^d$ is a function $\mathcal{F}\colon \overline{\Omega} \times \mathbb{R}^d \to \mathbb{R}$ of the form

$$\mathcal{F}_x(v) := \langle v, M(x)v\rangle^{1/2} + \langle \omega(x), v\rangle,$$

where M and ω are given fields, of respectively positive symmetric definite matrices and vectors, in Ω. The compatibility condition $\langle \omega(x), M(x)^{-1}\omega(x)\rangle < 1$ between both fields is assumed in Ω. In the particular case of an identically zero vector field ω, the Randers metric is reduced to a Riemannian metric.

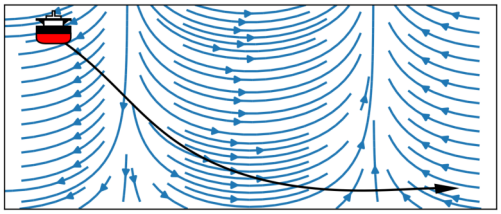


Figure 1.4: Zermelo's navigation problem. The aim is to compute the trajectory (in black) that allows the boat to reach its destination in minimal time, taking into account ocean currents (in blue). The minimal travel time between a given source and destination defines a Randers distance.

To two points x, $y \in \overline{\Omega}$ is associated the set Γ_x^y of all Lipschitz continuous paths $\gamma\colon [0,1] \to \overline{\Omega}$ between x and y (meaning that they satisfy $\gamma(0) = x$ and $\gamma(1) = y$). The length, with respect to a given Randers metric $\mathcal{F}$, of any such path is defined as follows:

$$\mathrm{length}_{\mathcal{F}}(\gamma) := \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t))\,dt.$$

The Randers distance between points x and y is defined as the minimal length among all paths:

$$\mathrm{dist}_{\mathcal{F}}(x,y) := \inf_{\gamma \in \Gamma_x^y} \mathrm{length}_{\mathcal{F}}(\gamma).$$

In Chapter 4, we introduce a numerical method for approximating the function $\mathbf{u}$ defined in $\overline{\Omega}$ by

$$\mathbf{u}(x) := \inf_{p \in \partial\Omega} (g(p) + \mathrm{dist}_{\mathcal{F}}(p,x)), \tag{1.11}$$

for some given $g\colon \partial\Omega \to \mathbb{R} \cup \{+\infty\}$ (up to excluding some point $p_0 \in \Omega$ from Ω, and letting $g(p_0) = 0$ and $g = +\infty$ on $\partial\Omega \setminus \{p_0\}$, this includes the particular case of the function $\mathbf{u}\colon x \mapsto \mathrm{dist}_{\mathcal{F}}(p_0,x)$). This method involves solving a linear finite difference scheme and is justified by a large deviations principle. Its advantages include, on the one hand, that the linearity of the scheme allows using prefactorization techniques for efficiently approximating distances between many pairs of points, and on the other hand, that it is particularly well suited to the numerical resolution of the entropic regularization of optimal transport problems. A similar method was previously introduced in the particular setting of Riemannian manifolds [CWW13], and was applied to the approximation of optimal transport distances in such manifolds [Sol+15].

The finite difference scheme that we recommend solving is of the form

$$\begin{aligned} &u_\varepsilon^h(x) + 2\varepsilon \sum_{i=1}^I \mu_i(x) \frac{u(x + he_i(x)) - u(x - he_i(x))}{2h} \\ &- \varepsilon^2 \sum_{i=1}^I \lambda_i(x) \frac{u(x + he_i(x)) + u(x - he_i(x)) - 2u(x)}{h^2} = 0 \quad \text{in } \Omega \cap h\mathbb{Z}^d, \end{aligned} \tag{1.12}$$

where coefficients μ_i, λ_i, and offsets e_i satisfy (1.3) and (1.9), b and $\mathcal{D}$ being fields in Ω of respectively vectors and positive symmetric defined matrices, defined from ω and M by simple algebraic relations. Note the similarity between the left-hand side in the scheme (1.12) and the

discretization (1.8). The scheme has to be adapted close to $\partial\Omega$ in order to take into account the Dirichlet boundary condition $u_\varepsilon^h(x) = \exp(-g(x)/\varepsilon)$ on $\partial\Omega$. The method states that the function $\mathbf{u}$ is approximated, under suitable assumptions, by $\mathbf{u}_\varepsilon^h := -\varepsilon \ln u_\varepsilon^h$, for small values of ε and h.

We display in Figure 1.5 some numerical results obtained by applying the proposed numerical method to Zermelo's navigation problem.

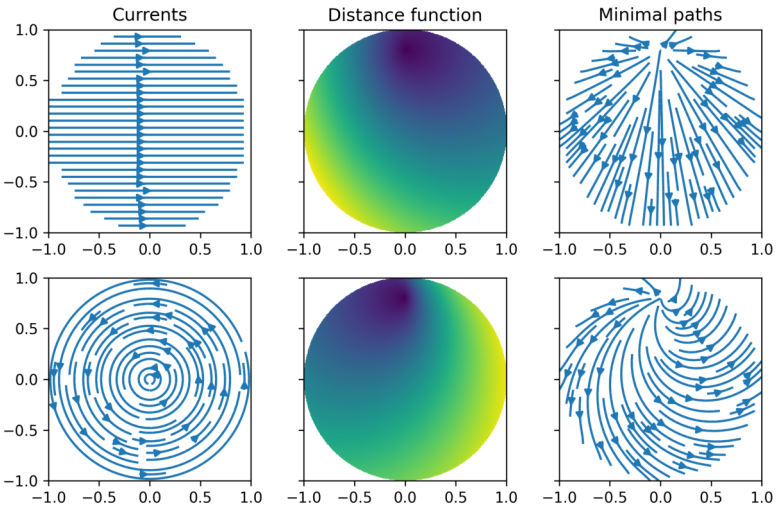


Figure 1.5: Application of the proposed numerical method to Zermelo's navigation problem (see Figure 1.4). Left: ambient currents. Middle: Randers distance function from the point $p_0 := (0, 0.8)$. Right: minimal paths from the point p_0.

Large deviations principle. It is well-known that the function $\mathbf{u}$ defined by (1.11) is solution to the Hamilton-Jacobi-Bellman equation

$$\langle D\mathbf{u}(x), \mathcal{D}(x) D\mathbf{u}(x)\rangle + 2\langle b(x), D\mathbf{u}(x)\rangle - 1 = 0 \quad \text{in } \Omega, \tag{1.13}$$

where $\mathcal{D}$ and b are the fields introduced above. On the other hand, the scheme (1.12) is a discretization of the linear second-order equation

$$u_\varepsilon(x) + 2\varepsilon\langle b(x), Du_\varepsilon(x)\rangle - \varepsilon^2 \operatorname{Tr}(\mathcal{D}(x) D^2 u_\varepsilon(x)) = 0 \quad \text{in } \Omega. \tag{1.14}$$

Formally, it is easy to show that if u_ε is solution to (1.14), then $\mathbf{u}_\varepsilon := -\varepsilon \ln u_\varepsilon$ is solution to

$$\langle D\mathbf{u}_\varepsilon(x), \mathcal{D}(x) D\mathbf{u}_\varepsilon(x)\rangle + 2\langle b(x), D\mathbf{u}_\varepsilon(x)\rangle - \varepsilon \operatorname{Tr}(\mathcal{D}(x) D^2 \mathbf{u}_\varepsilon(x)) - 1 = 0 \quad \text{in } \Omega. \tag{1.15}$$

The Hamilton-Jacobi-Bellman equation (1.15) is a perturbation of (1.13), justifying the convergence of $\mathbf{u}_\varepsilon$ to $\mathbf{u}$ as ε approaches zero.

While the above justification is formal, we explain in section 4.A how to make it rigorous in the setting of viscosity solutions.

We call the result $\mathbf{u}_\varepsilon \to_{\varepsilon\to 0} \mathbf{u}$ a large deviations principle since the function $\mathbf{u}_\varepsilon$ has the following probabilistic interpretation: for given $x \in \overline{\Omega}$ and $\varepsilon > 0$, let $(X_t^{x,\varepsilon})_{t\geq 0}$ be the stochastic process defined by

$$dX_t^{x,\varepsilon} = -2\varepsilon b(X_t^{x,\varepsilon})\, dt + \varepsilon\sqrt{2\mathcal{D}}(X_t^{x,\varepsilon})\, dW_t, \qquad X_0^{x,\varepsilon} = x,$$

where $(W_t)_{t\geq 0}$ is a d-dimensional Wiener process, and let $\tau^{x,\varepsilon} \geq 0$ be the exit time

$$\tau^{x,\varepsilon} := \inf\{t \geq 0 \mid X_t^{x,\varepsilon} \notin \Omega\}.$$

Then $\mathbf{u}_\varepsilon(x)$ is equal to

$$-\varepsilon \ln\left(\mathbb{E}\left[\exp\left(\frac{-\varepsilon\tau^{x,\varepsilon} - g(X_{\tau^{x,\varepsilon}}^{x,\varepsilon})}{\varepsilon}\right)\right]\right),$$

which may be interpreted as a soft-minimum, and compared to the infimum in (1.11).

Convergence analysis. In addition to the convergence $\mathbf{u}_\varepsilon \to_{\varepsilon\to 0} \mathbf{u}$, under suitable assumptions it is easily proved, using the standard theory of convergence of finite difference schemes for linear equations and the fact that the scheme (1.12) is consistent with the equation (1.14), that $u_\varepsilon^h \to_{h\to 0} u_\varepsilon$, and thus also that $\mathbf{u}_\varepsilon^h \to_{h\to 0} \mathbf{u}_\varepsilon$. However, this does not guarantee the joint convergence $\mathbf{u}_\varepsilon^h \to_{(\varepsilon,h)\to 0} \mathbf{u}$.

We study the joint convergence in two different settings. In the first one, that we call here the *smooth setting*, we assume, among other assumptions, that the domain Ω is smooth and that the function g in (1.11) is continuous and takes finite values. In the second one, that we call the *singular setting*, we assume that the boundary of Ω is the union of a smooth part and of some isolated point p_0, and we let $g(p_0) = 0$ and $g = +\infty$ elsewhere: as already discussed above, this allows us to consider the case of the function $\mathbf{u}\colon x \mapsto \mathrm{dist}_{\mathcal{F}}(p_0, x)$, which is often the one that has to be approximated in practical applications.

In the smooth setting, our strategy is, instead of performing a logarithmic transformation in the linear equation (1.14) in order to obtain the nonlinear equation (1.15), to rather perform the logarithmic transformation directly in the linear finite difference scheme (1.12), which yields a nonlinear, monotone scheme to which $\mathbf{u}_\varepsilon^h := -\varepsilon \ln u_\varepsilon^h$ is solution. Under suitable assumptions, we show (Proposition 4.3.13) the consistency of this nonlinear scheme with the equation (1.13) as $\varepsilon \to 0$ and $h/\varepsilon \to 0$, and we deduce (Theorem 4.3.18) the convergence of $\mathbf{u}_\varepsilon^h$ towards $\mathbf{u}$. Importantly, the convergence result does not hold anymore if the ratio h/ε remains constant, which is reminiscent of the counterexamples to the convergence of the method in [CWW13] towards the distance function, see [CWW13, Appendix A]. We recommend choosing the parameter ε proportionally to $h^{2/3}$. We show (Corollary 4.3.14) that, at least far from $\partial\Omega$, this yields consistency of the nonlinear scheme at the order 2/3, which is the best achievable order of consistency among all possible choices of the parameter ε. For comparison, we show that if upwind finite differences had been used instead of centered finite differences for the discretization of the first-order term in the scheme (1.12), then only consistency at the order 1/2 could have been achieved for the nonlinear scheme.

In the singular setting, we prove under suitable assumptions (Theorem 4.4.1) the convergence of $\mathbf{u}_\varepsilon^h$ towards $\mathbf{u}$ as $\varepsilon \to 0$, $h/\varepsilon \to 0$, and $\varepsilon \ln h \to 0$. To this end, we reuse when appropriate the arguments from the smooth setting, but in order to handle the isolated point in the boundary of Ω, we also need to use a two- or three-dimensional counterpart to the *spanning property* discussed in section 1.1.2, which prevents the formation of checkerboard artifacts close to this point.

Finally, we discuss the approximation of the Randers distance $\mathrm{dist}_{\mathcal{F}}(x, y)$ when neither of the points x and y is fixed. The scheme (1.12), modified close to $\partial\Omega$ to take into account the Dirichlet boundary condition $u_\varepsilon^h = 0$ on $\partial\Omega$, may be written in the matrix form $L_\varepsilon^h u_\varepsilon^h = 0$, where L_ε^h is a square matrix indexed by x, $y \in \Omega \cap h\mathbb{Z}^d$. We show (Theorem 4.4.2) that, under suitable assumptions and locally uniformly over $(x, y) \in \Omega \times \Omega$,

$$-\varepsilon \ln[((L_\varepsilon^h)^{-1})_{xy}] \to \mathrm{dist}_{\mathcal{F}}(x, y), \qquad (\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0. \tag{1.16}$$

When approximating distances $\mathrm{dist}_{\mathcal{F}}(x,y)$ between many pairs of points $(x,y) \in \Omega \times \Omega$, we recommend prefactorizing the matrix L_ε^h so that the elements $((L_\varepsilon^h)^{-1})_{xy}$ may be computed efficiently.

Application to regularized optimal transport. Given two probability measures μ and ν supported on $\Omega \cap h\mathbb{Z}^d$, the 1-Wasserstein optimal transport problem between μ and ν writes as

$$W(\mu,\nu) := \inf_{P \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \mathrm{dist}_{\mathcal{F}}(x,y)\, dP(x,y),$$

where $\Pi(\mu,\nu)$ denotes the set of *transport plans* between μ and ν, that is, the set of probability measures on $\Omega \times \Omega$ whose first and second marginals coincide respectively with μ and ν. This problem admits the entropic regularization

$$W_\varepsilon(\mu,\nu) := \inf_{P \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \mathrm{dist}_{\mathcal{F}}(x,y)\, dP(x,y) - \varepsilon\, \mathrm{Ent}(P),$$

where $\mathrm{Ent}(P) := -\sum_{x,y\in\Omega\cap h\mathbb{Z}^d} P_{xy} \ln P_{xy}$ if $P = \sum_{x,y\in\Omega\cap h\mathbb{Z}^d} P_{xy}\delta_{(x,y)}$.

It is well-known that the entropic regularization of the optimal transport problem may be solved using *Sinkhorn's algorithm* [Cut13]. This algorithm involves performing several matrix-vector products involving the square matrix K_ε indexed by x, $y \in \Omega \cap h\mathbb{Z}^d$ and defined by

$$(K_\varepsilon)_{xy} := \exp\left(-\frac{\mathrm{dist}_{\mathcal{F}}(x,y)}{\varepsilon}\right).$$

Since the matrix K_ε is dense, it would be numerically extremely costly to compute it explicitly, and thus it is desirable to find instead an approximation of K_ε for which the matrix-vector products may be computed efficiently. According to (1.16), we recommend approximating K_ε by $(L_\varepsilon^h)^{-1}$. Then matrix-vector products involving $(L_\varepsilon^h)^{-1}$ reduce to sparse linear systems, which may be solved efficiently especially if the matrix L_ε^h has been prefactorized beforehand.

A similar approach was previously introduced in [Sol+15] in the setting of Riemannian manifolds. We notice that while optimal transport problems in which the transport cost $\mathrm{dist}_{\mathcal{F}}(x,y)$ has been replaced by its square $\mathrm{dist}_{\mathcal{F}}(x,y)^2$ were also discussed in [Sol+15], the approach used to handle this quadratic variant cannot be generalized to the setting of Randers manifolds.

We display in Figure 1.6 some numerical results obtained by solving optimal transport problems using the proposed method, in Randers manifold associated to Zermelo's navigation problem.

1.2.2 Monotone and second order consistent scheme for the two-dimensional Pucci equation

In Chapter 5, we introduce a monotone discretization of the two-dimensional Pucci equation

$$\lambda_{\min}(D^2u(x)) + \mu\lambda_{\max}(D^2u(x)) = f(x), \tag{1.17}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ respectively denote the smallest and largest eigenvalues, and $\mu > 0$ is a given parameter. For any $\theta \in \mathbb{R}$, we define the rotation matrix R_θ and the rotated matrix $\mathcal{D}(\theta,\mu)$ by

$$R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \mathcal{D}(\theta,\mu) := R_\theta \begin{pmatrix} 1 & 0 \\ 0 & \mu \end{pmatrix} R_\theta^\top;$$

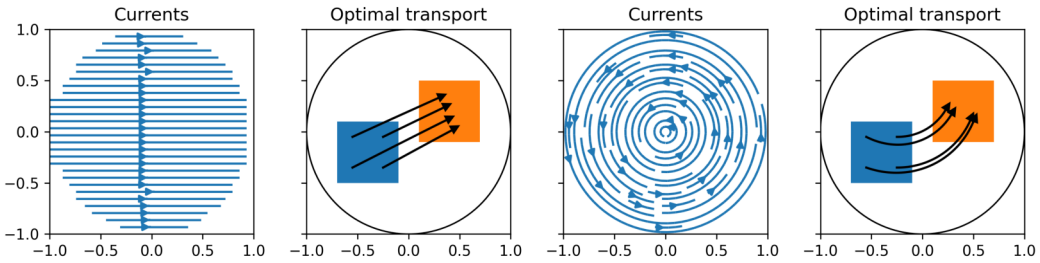


Figure 1.6: Optimal transport in the setting of Zermelo's navigation problem. Left, middle right: ambient currents. Middle left, right: each point in the source domain (in blue) is mapped by the optimal transport plan to a measure in the target domain (in orange). We display arrows pointing to the barycenters of those measures. The fact that the optimal transport is not a translation even when the currents are constant is a well-known property of the 1-Wasserstein problem.

then the Pucci equation (1.17) admits the following reformulation in the form (1.2):

$$\max_{\theta\in[0,\pi]} -\operatorname{Tr}(\mathcal{D}(\theta,\mu)D^2u(x)) = -f(x), \tag{1.18}$$

provided that $\mu \leq 1$ (if $\mu \geq 1$, then the maximum has to be replaced by a minimum). Following the approach described in section 1.1 in order to discretize the reformulation (1.18) yields the finite difference scheme

$$\max_{\theta\in[0,\pi]} -\sum_{e\in\mathbb{Z}^d\setminus\{0\}} \lambda^e(\theta,\mu)\frac{u(x+he)+u(x-he)-2u(x)}{h^2} = -f(x), \tag{1.19}$$

where, for a fixed value of $\mu > 0$, the coefficients $\lambda^e(\theta,\mu) \geq 0$ are nonzero only for finitely many offsets $e \in \mathbb{Z}^d \setminus \{0\}$ and may be computed using *Selling's algorithm*, see section 1.1.1.

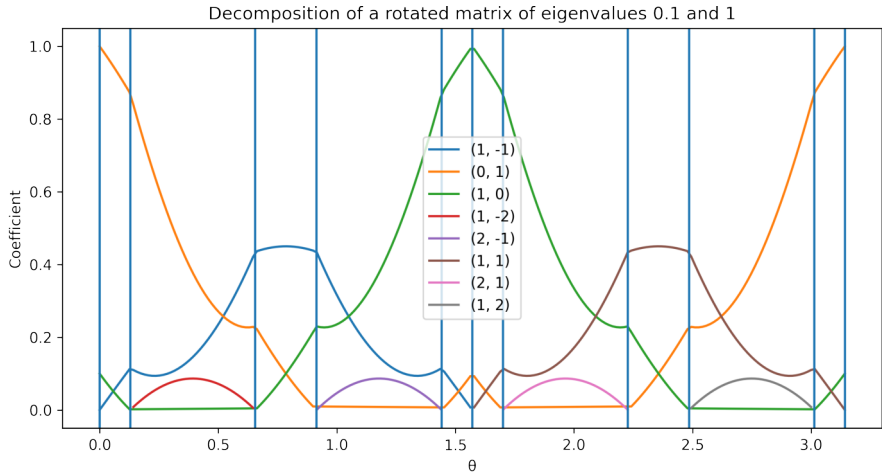


Figure 1.7: Coefficients $\lambda^e(\theta, 0.1)$, for different values of the offset e and of the angle θ.

A common approach to evaluate the left-hand side in the scheme (1.19) is to discretize the parameter set $[0,\pi]$ of the maximum, which introduces an approximation error. Then, for each θ

in the discretized parameter set, the nonzero coefficients $\lambda^e(\theta,\mu)$, and the corresponding offsets, may be computed using Selling's algorithm. This approach is numerically costly: if the Cartesian grid on which the scheme (1.19) is posed has a number of elements of the order of N^2, then the number of elements of the discretization of the parameter set should also be of the order of N^2 in order for the scheme to remain consistent at the order two, and the numerical cost of evaluating the left-hand side in (1.19) at all points of the Cartesian grid would be of the order of N^4.

Our main contribution is to show (section 5.2.2) that the maximum in (1.19) admits a closed-form formula. Therefore the scheme may be evaluated without discretizing the parameter set, and the numerical cost of this evaluation on the above Cartesian grid is only of the order of N^2. In order to compute the closed-form formula, we show that, as illustrated in Figure 1.7, functions $\theta \mapsto \lambda^e(\theta,\mu)$ coincide with the sum of a sinusoid and a constant on each one of finitely many closed intervals whose union is $[0,\pi]$. Each of those intervals $I \subset [0,\pi]$ is associated to a set of symmetric positive definite matrices with normalized trace

$$\left\{(1+\mu)^{-1}\mathcal{D}(\theta,\mu) \mid \theta \in I\right\}$$

which is included in a single cell of the infinite triangulation displayed in Figure 1.2.

We conclude Chapter 5 by performing some numerical experiments which illustrate the accuracy of the recommended numerical method.

While we apply to the Pucci equation the idea of computing a closed-form formula for the maximum occurring in a finite difference scheme designed using the tools from section 1.1, this approach can be generalized to other degenerate elliptic partial differential equations. We discuss the case of the Monge-Ampère equation in section 1.2.3.

1.2.3 Monotone discretization of the Monge-Ampère equation of optimal transport

In Chapter 6, we introduce a monotone finite difference discretization of the second boundary value problem for the Monge-Ampère equation. This is the relevant boundary value problem for Monge-Ampère equations associated to optimal transport problems. We consider Monge-Ampère equations of the form

$$\det\left(D^2u(x) - A(x,Du(x))\right) = B(x,Du(x)), \tag{1.20}$$

where $B \geq 0$ and matrices $A(x,Du(x))$ are symmetric, our convergence result only being proved in the particular case of equations of the form

$$\det D^2u(x) = \frac{f(x)}{g(Du(x))}, \tag{1.21}$$

which corresponds to optimal transport problems whose cost function is the squared Euclidean distance (we call those *quadratic* optimal transport problems). Solutions to the Monge-Ampère equation are considered admissible when they are convex, in the setting of equation (1.21), or when they satisfy formally the generalized convexity condition $D^2u(x) \succeq A(x,Du(x))$ for the Loewner order on symmetric matrices, in the setting of equation (1.20).

Discretization. While the general form of the numerical scheme that we recommend is described in Chapter 6, let us discuss here, for simplicity, the discretization of the simpler equation

$$\det D^2u(x) = f(x).$$

This equation admits the following reformulation in the form (1.2), which was previously used in [FJ17] for numerical purposes:

$$\max_{\mathcal{D}\in\mathbf{S}_1} \Big(df(x)^{1/d} (\det \mathcal{D})^{1/d} - \operatorname{Tr}(\mathcal{D} D^2 u(x)) \Big) = 0, \tag{1.22}$$

where d is the dimension of the domain of the equation and $\mathbf{S}_1$ denotes the space of symmetric positive semidefinite matrices of size d whose trace is equal to one. The proposed scheme is a finite difference discretization of the above reformulation:

$$F^h_{\mathrm{MA}} u(x) = 0, \qquad F^h_{\mathrm{MA}} u(x) := \max_{\mathcal{D}\in\mathbf{S}_1^h} \Big(df(x)^{1/d} (\det \mathcal{D})^{1/d} - \Delta_h^{\mathcal{D}} u(x) \Big), \tag{1.23}$$

where $\mathbf{S}_1^h$ is a suitable approximation of $\mathbf{S}_1$, and $\Delta_h^{\mathcal{D}} u(x)$ is a finite difference discretization of $\operatorname{Tr}(\mathcal{D} D^2 u(x))$ obtained following (1.4), with an appropriate choice of coefficients $\lambda_i \geq 0$ and offsets $e_i \in \mathbb{Z}^d$ satisfying (1.3).

One benefit of using the reformulation (1.22) of the Monge-Ampère equation is that the equation (1.22) enforces the convexity of its solutions. As a consequence, large steps may be used when solving the associated finite difference scheme using Newton's method. For comparison, when solving some previously introduced finite difference schemes for the Monge-Ampère equation, such as [BCM16], extremely small steps needed to be used in Newton's method in order to preserve the convexity of the iterates.

In dimension $d = 2$, we recommend choosing the coefficients λ_i and the offsets e_i in $\Delta_h^{\mathcal{D}} u(x)$ as *Selling's decomposition* of the matrix $\mathcal{D}$, consistently with the approach described in section 1.1. We also recommend choosing the set $\mathbf{S}_1^h \approx \mathbf{S}_1$ as a finite subtriangulation of the infinite triangulation of the disk $\mathbf{S}_1$ displayed in Figure 1.2. As in the case of the Pucci equation (see section 1.2.2), we prove (Theorem 6.1.2) that the maximum in (1.23) then admits a closed-form formula, the proof using the fact that offsets e_i associated to nonzero coefficients λ_i remain constant on each cell of the triangulation $\mathbf{S}_1^h$. Thanks to this closed-form formula, the discrete operator F^h_{MA} in (1.23) may be evaluated particularly efficiently.

	General case	Smooth case, with Lax-Friedrichs	Smooth case, without Lax-Friedrichs
Consistency error	$O(h^{2/3})$	$O(h)$	$O(h^2)$
Numerical cost	$O(h^{-8/3}\log(1+h^{-1}))$	$O(h^{-2})$	$O(h^{-2})$
Numerical cost (discretized maximum)	$O(h^{-10/3})$	$O(h^{-6})$	$O(h^{-4})$

Table 1.3: Analysis of the consistency error and numerical cost of the proposed finite difference discretization of the Monge-Ampère operator.

In Table 1.3, we display the order of consistency of the scheme (1.23) with the reformulated Monge-Ampère equation (1.22), and the numerical cost of evaluating the operator F^h_{MA}. For comparison, we also display an optimistic estimation of the numerical cost of evaluating the operator F^h_{MA} when discretizing the parameter set of the maximum instead of using the closed-form formula, and when the discretization of the parameter set has to be fine enough so that the order of consistency of the scheme has to be preserved. Our analysis also applies to the generalization of the scheme (1.23) to Monge-Ampère equations of the general form (1.20). We distinguish between three different cases. In the *general case*, a part of the consistency error is due to the approximation of the parameter set $\mathbf{S}_1$ of the maximum by $\mathbf{S}_1^h$, and the number of

cells in the triangulation $\mathbf{S}_1^h$ has to increase as the discretization step h decreases, which has an effect to the numerical cost of the method. In the *smooth case*, we assume that the solution to the Monge-Ampère equation is smooth and satisfies a strict uniform variant of the generalized convexity condition $D^2u(x) \succeq A(x, Du(x))$: then it is possible to choose the set $\mathbf{S}_1^h$ independently of h, which both increases the order of consistency of the scheme and decreases the numerical cost. When discretizing Monge-Ampère equations of the form (1.20) for which either function A or B effectively depend on its second variable, we recommend using a Lax-Friedrichs approximation of the gradient $Du(x)$, which is only consistent to the order one; thus we have to distinguish two subcases of the smooth case, according to whether such a Lax-Friedrichs approximation is needed or not.

The optimal transport boundary condition. We assume from now on that the Monge-Ampère equation is posed on a bounded domain $X \subset \mathbb{R}^d$. In the setting (1.21) of Monge-Ampère equations associated to quadratic optimal transport problems, the second boundary value problem for the Monge-Ampère equation involves the optimal transport boundary condition

$$\overline{Du(X)} = \overline{Y}, \tag{1.24}$$

where Y is some given convex set. The above equality is called a boundary condition even though it involves the whole domain X: this is because under suitable assumptions, by a convexity argument, it may be reformulated as $Du(\partial X) = \partial Y$.

The second boundary value problem for the Monge-Ampère equation features a property of *additive invariance*: both the Monge-Ampère equation and the optimal transport boundary condition involve only the derivatives of the unknown u, and not its values. For this reason, the set of solutions to the boundary value problem is stable by addition of a constant.

The property of additive invariance is a source of difficulty when designing numerical schemes. In particular, it may prevent the existence of solutions to those schemes. Let us illustrate this phenomenon with the example of another additively invariant system of equations:

$$\begin{cases} -u''(x) + f(x) = 0 & \text{in } (-1,1), \\ u'(-1) = u'(1) = 0. \end{cases} \tag{1.25}$$

This is Poisson's equation on the one-dimensional domain $(-1,1)$, equipped with Neumann boundary conditions. It only admits solutions provided that the following condition of compatibility between the source term f and the boundary conditions is satisfied:

$$\int_{-1}^{1} f(x)\,dx = 0. \tag{1.26}$$

A similar compatibility condition has to be satisfied in order for the Monge-Ampère system to admit solutions, in form of the mass balance condition

$$\int_X f(x)\,dx = \int_Y g(y)\,dy. \tag{1.27}$$

The difficulty, in the setting of finite difference schemes, is that often no discrete counterpart to (1.26) or (1.27) is satisfied. A natural finite difference scheme for the system (1.25) on the grid $(-1,1) \cap h\mathbb{Z}$ would be

$$-\frac{\delta_h^+ u(x) + \delta_h^- u(x)}{h} + f(x) = 0 \quad \text{in } (-1,1) \cap h\mathbb{Z},$$

$$\delta_h^{\pm} u(x) := \begin{cases} \frac{u(x\pm h)-u(x)}{h} & \text{if } u(x \pm h) \in (-1,1), \\ 0 & \text{else.} \end{cases}$$

This scheme does not admit solutions, except for the special case of source terms f satisfying the discrete compatibility condition

$$\sum_{x \in (-1,1) \cap h\mathbb{Z}} f(x) = 0.$$

While the above scheme may be adapted in order to admit solutions, not all the possible adaptations may be generalized to finite difference discretizations of the Monge-Ampère equation.

The discretization that we introduce for the optimal transport boundary condition (1.24) is based on the fact that the inclusion $Du(X) \subset \overline{Y}$ may be rewritten as the inequality

$$\max_{|e|=1} (\langle e, Du(x) \rangle - \sigma_Y(e)) \leq 0 \quad \text{in } X,$$

where $\sigma_Y : e \mapsto \sup_{p \in P(x)} \langle e, p \rangle$ is the *convex support function* of the set $P(x)$. This reformulation may be discretized using standard upwind finite differences, yielding the scheme

$$F^h_{\mathrm{BV2}} u(x) \leq 0 \qquad\qquad F^h_{\mathrm{BV2}} u(x) := \max_{|e|=1} (D^e_h u(x) - \sigma_{P(x)}(e)). \tag{1.28}$$

We now have an equality (1.23) and an inequality (1.28), both posed on a Cartesian discretization of the domain X. Our aim is to design a numerical scheme for the whole system of equations of the second boundary value problem, using both discrete operators F^h_{MA} and F^h_{BV2}.

One way to do this is to adapt to our setting the approach developed in [Fro19]. This would yield the scheme

$$\max\{F^h_{\mathrm{MA}} u(x), F^h_{\mathrm{BV2}} u(x)\} = 0, \tag{1.29}$$

modified close to ∂X to take into account the Dirichlet boundary condition $u = 0$ on ∂X. This approach was studied in [Fro19] in the setting (1.21) of quadratic optimal transport problems, although not with the same definitions of the operators F^h_{MA} and F^h_{BV2}. Taking the maximum between operators F^h_{MA} and F^h_{BV2} was justified by a phenomenon of competition between the inequalities $F^h_{\mathrm{MA}} u(x) \leq 0$ and $F^h_{\mathrm{BV2}} u(x) \leq 0$. The Dirichlet boundary condition is satisfied in the classical sense $u(x_*) = 0$ close to some point $x_* \in \partial X$, while at other points it is to be understood in the weak sense $u \leq 0$. It serves two purposes: on the one hand, the equality $u(x_*) = 0$ selects a unique solution in the additively invariant set of solutions to the Monge-Ampère problem, and on the other hand, it weakens the discretization of the optimal transport boundary condition close to the point x_*, thus also weakening the need for a discrete counterpart to the mass balance condition (1.27). A key ingredient in the analysis in [Fro19] is that the numerical scheme satisfies a property of *underestimation*. One difficulty when attempting to extend this analysis to Monge-Ampère equations of the general form (1.20) is that it does not seem obvious how to discretize (1.20) in an *underestimating* manner.

In order to bypass the need for the underestimating property, and also in order to prevent numerical artifacts that tend to occur close to the point x_*, we use a slight modification of the scheme (1.29). Similarly to the numerical experiments in [BD19], we add an unknown $\alpha \in \mathbb{R}$ to the discrete problem, and we solve

$$\max\{F^h_{\mathrm{MA}} u(x) + \alpha, F^h_{\mathrm{BV2}} u(x)\} = 0. \tag{1.30}$$

The presence of the unknown α is sufficient to weaken the need for a discrete counterpart to the mass balance condition (1.27). We replace the boundary condition $u = 0$ by $u = +\infty$ on ∂X.

The scheme (1.30) only enforces this Dirichlet boundary condition in the weak sense $u \leq +\infty$ — which is always satisfied — on the whole boundary ∂X, and we show that the Dirichlet boundary condition yields no boundary layer. In order to select a unique solution in the additively invariant set of solutions to the Monge-Ampère problem, we add a constraint $u(x_0) = 0$, for some given point $x_0 \in X$. Adding an equality constraint is consistent with the fact of adding an unknown α to the scheme.

One drawback of adding the unknown α is that, while operators F^h_{MA} and F^h_{BV2} are both monotone with respect to their argument u, the scheme (1.30) is not monotone with respect to the pair of unknowns (u, α). Therefore Perron's method, which often allows to prove the existence of solutions to monotone numerical schemes, does not apply directly to the scheme (1.30). We show (Theorem 6.2.14) that Perron's method may however be adapted to our setting by handling the unknown α separately from u in the proof of the existence of solutions. We state our proof as a general result about the existence of solutions to a class of additively invariant numerical schemes whose unknowns are a discrete function u and some scalar $\alpha \in \mathbb{R}$ and which are monotone with respect to u for fixed values of α.

Under suitable assumptions, we establish the convergence of the numerical scheme (1.30) in the setting (1.21) of quadratic optimal transport problems (Theorem 6.5.22). To this end, we need to study the relationship between, on the one hand, *viscosity* subsolutions and supersolutions to the discretized system of equations, and on the other hand, *Aleksandrov* solutions to the Monge-Ampère problem. The case of viscosity subsolutions was previously studied in the setting of [Fro19], but not the case of viscosity supersolutions, whose analysis was not needed in that setting.

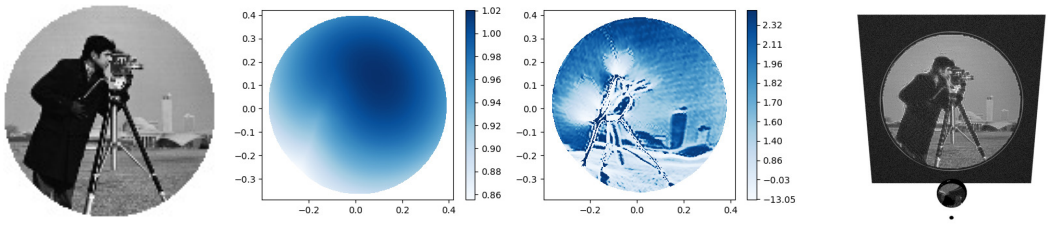


Figure 1.8: Application of the method to the far field refractor problem in nonimaging optics. Left: target image. Middle left: shape of the lens, computed using an implementation of the method in the Python® programming language. Middle right: approximation of the pointwise curvature of the lens. Right: simulation of the scene, using the appleseed® rendering engine. The small black dot at the bottom represents the light source.

While the convergence of the scheme (1.30) remains an open problem when considering Monge-Ampère equations of the general form (1.20), we performed numerical experiments validating the method in the setting of the far field refractor problem in nonimaging optics. In this problem, rays of light emanate from a point light source, are refracted by an optical lens, and the refracted rays continue to propagate until they hit a screen, was distance to the point source and to the lens is assumed to be large. The aim is to find a suitable shape for the lens so that a given target image is projected to the screen. It is well-known [GH09] that this problem is described by some Monge-Ampère equation of the form (1.20). We solve the associated numerical scheme (1.30), we deduce an approximation of the shape of the lens, and we simulate the propagation of the light for the shape that we obtain, as illustrated in Figure 1.8. We observe that graph of the approximated pointwise curvature of the lens is reminiscent of the target image, but that bright

areas of the target image tend to be expanded and that dark areas tend to be shrunk, which is an expected result since parts of the lens corresponding to bright areas must be large in order to capture more rays of light.

Part I

Tools for the monotone finite difference discretization of anisotropic differential operators on Cartesian grids

Chapter 2

Second order monotone finite differences discretization of linear anisotropic differential operators

This chapter corresponds to the paper [BBM21c].

2.1 Introduction

In this paper, we design finite difference discretizations of Degenerate Elliptic (DE) Partial Differential Equations (PDEs). This class of equations is sufficiently general to encompass a wide variety of applications, in the fields of optimal transport, game theory, differential geometry, stochastic modeling and finance, optimal control, etc. Our results are limited to linear and semi-linear operators, but could in principle be used as a building block for the discretization of fully non-linear operators, see Appendix 2.A. The assumption of degenerate ellipticity yields comparison principles and stability properties [CIL92].

Discrete Degenerate Ellipticity (DDE), for numerical schemes, implies similarly strong properties [Obe06], which often turn proofs of convergence into simple verifications. A known limitation of DDE discretization schemes is their consistency order with the original PDE, which cannot exceed two for second order operators and one for first order operators [Obe06]. However, many common implementations of second order DE operators only achieve first order consistency, or sometimes less. They may also rely on excessively wide stencils, especially in the context of two-scales discretizations [FO11; LN18]. This degrades the accuracy of the numerical results, which severely constrains the practical uses of these methods. The objective of this paper is to characterize when a second order monotone discretization is feasible, and how wide the numerical scheme stencil must be, especially when the second order part of the operator is strongly anisotropic, and the first order term is non-vanishing.

The finite difference schemes developed in this paper are adaptive in the sense that the stencil of the numerical scheme depends on the PDE coefficients, and reflects the strength and orientation of the anisotropy of the PDE, see Figure 2.1. We do however use a fixed and non-adaptive Cartesian discretization grid $\Omega \cap h\mathbb{Z}^d$. This understanding of adaptivity must be distinguished from the (more standard) approach where the sampling density of the set of discretization points is adjusted locally in the PDE domain, often depending on the singularities of the solution and using a priori or a posteriori estimators, see e.g. [DK09]. In other words, we do not in this work

adapt the *set* of discretization points to the addressed PDE and its solution, but we adapt the *connectivity* between these points as defined by the numerical scheme stencils. This paper builds on [Mir16] which similarly discusses optimally compact stencils for finite differences discretizations of several PDEs—divergence form and non-divergence form anisotropic diffusion, anisotropic eikonal equation—preserving suitable structural properties, in two dimensions. Regarding the non-divergence form Laplacian, the present work differs from [Mir16] by focusing on the obstructions to discretization related to the presence of an additional first order term, and by addressing the three-dimensional case.

We state our theoretical results in the context of linear operators with constant coefficients, defined over $\mathbb{R}^d$ where $d \in \{2, 3\}$. Because degenerate ellipticity is a local property, which is stable under a variety of transformations, they admit straightforward extensions to semi-linear operators and to some fully non-linear operators. Non-constant coefficients and bounded domains with Dirichlet boundary conditions are also easily handled. See Appendix 2.A and the numerical experiments section 2.4 for these extensions.

We define the linear operator $\mathcal{L} = \mathcal{L}[\omega, D]$ on $\mathbb{R}^d$ by the expression

$$-\mathcal{L}u(x) := \langle \omega, \nabla u(x) \rangle + \tfrac{1}{2} \operatorname{Tr}(D\nabla^2 u(x)), \tag{2.1}$$

where $\omega \in \mathbb{R}^d$, $D \in S_d^{++}$ is a symmetric positive definite matrix, and the unknown $u : \mathbb{R}^d \to \mathbb{R}$ is a smooth function. Likewise in the discrete setting we define the finite difference operator $L_h = L_h[\rho_i, e_i]_{1 \le |i| \le I}$, on the Cartesian grid $h\mathbb{Z}^d$ with grid scale $h > 0$, by the expression

$$-L_h u(x) := h^{-2} \sum_{1 \le |i| \le I} \rho_i \big(u(x + he_i) - u(x)\big), \tag{2.2}$$

where $\rho_{-I}, \cdots, \rho_{-1}, \rho_1, \cdots, \rho_I \ge 0$ are non-negative weights, and $e_1, \cdots, e_i \in \mathbb{Z}^d$ are offsets with integer entries, for some positive integer I. Here and throughout this paper, without loss of generality, we use the convention that $e_{-i} := -e_i$ for all $1 \le i \le I$. Note that consistency between (2.1) and (2.2) across grid scales $h > 0$ requires that the weights $(\rho_i)_{1 \le |i| \le I}$ depend on h, in addition to ω and D.

Any translation invariant linear operator on $h\mathbb{Z}^d$, finitely supported and vanishing on constant functions, can be written in the form (2.2). We denote by S_d the set of symmetric $d \times d$ matrices, by S_d^+ the subset of semi-definite ones, and by S_d^{++} the positive definite ones.

Definition 2.1.1. The operator $\mathcal{L}[\omega, D]$ is said Degenerate Elliptic (DE) if $D \in S_d^+$. The discrete operator $L_h[\rho_i, e_i]_{1 \le |i| \le I}$ is said Discrete Degenerate Elliptic (DDE) if $\rho_i \ge 0$ for all $1 \le |i| \le I$.

In particular, the DE property does not impose any restrictions on the first order term $\omega \in \mathbb{R}^d$, and the DDE property does not constrain the numerical scheme offsets $e_i \in \mathbb{Z}^d$. The objective of this paper is to construct second order accurate DDE discretizations of DE operators, and to investigate the possible obstructions to do so. For that purpose we introduce the following compatibility condition.

Definition 2.1.2 (Absolute feasibility)**.** Let $\omega \in \mathbb{R}^d$ and $D \in S_d^+$. We say that the pair (ω, D) is *absolutely feasible* if there exists an integer $I \ge 0$, some integral offsets $e_1, \cdots, e_I \in \mathbb{Z}^d \setminus \{0\}$, and some non-negative weights $\rho_{-I}, \cdots, \rho_{-1}, \rho_1 \cdots, \rho_I \ge 0$, such that denoting $e_{-i} := -e_i$ for all $1 \le i \le I$ one has

$$\sum_{1 \le |i| \le I} \rho_i e_i = \omega, \qquad \sum_{1 \le |i| \le I} \rho_i e_i e_i^\top = D. \tag{2.3}$$

Let us emphasize that in Definition 2.1.2 the integer I and the offsets $(e_i)_{1\leq|i|\leq I}$ are not fixed a priori (and neither are the weights $(\rho_i)_{1\leq|i|\leq I}$), but they may depend on ω and D; when they exist, they are in general not unique. In this paper, we fully characterize when a pair (ω, D) is absolutely feasible in dimension $d \in \{2, 3\}$, see Proposition 2.3.9, which is not straightforward unless D is a diagonal matrix. We also advocate for a specific construction, see Definition 2.1.5, for which the stencil cardinality I is bounded in terms of the dimension d, whereas the stencil width $\max\{\|e_i\|;\, 1 \leq i \leq I\}$ is bounded in terms of condition number of D, see Theorem 2.1.9 below.

Proposition 2.1.3. *The pair $(h\omega, D)$ is* absolutely feasible *iff there exists $(\rho_i, e_i)_{1\leq|i|\leq I}$ such that $\mathcal{L}[\omega, D]$ and $L_h[\rho_i, e_i]_{1\leq|i|\leq I}$ are both degenerate elliptic, and are equal on all quadratic functions.*

Proof. The scheme $L_h[\rho_i, e_i]_{1\leq|i|\leq I}$ is well-defined iff the vectors e_i have integer coordinates, and is DDE iff the weights ρ_i are non-negative, the same conditions that arise in Definition 2.1.2. For any *quadratic* function $u : \mathbb{R}^d \to \mathbb{R}$, any $e \in \mathbb{Z}^d$ and any $h > 0$ one has the *exact* Taylor expansion

$$u(x + he) = u(x) + h\langle p, e\rangle + \frac{1}{2}h^2\langle e, Me\rangle,$$

with $p := \nabla u(x)$ and $M := \nabla^2 u(x)$. Therefore, using that $\langle e, Me\rangle = \operatorname{Tr}(Mee^T)$,

$$\begin{aligned} L_h u(x) &= \frac{1}{h^2}\sum_{1\leq|i|\leq I} \rho_i\big(h\langle p, e_i\rangle + \frac{1}{2}h^2 \operatorname{Tr}(Me_ie_i^T)\big) \\ &= \frac{1}{h}\Big\langle p, \sum_{1\leq|i|\leq I} \rho_i e_i\Big\rangle + \frac{1}{2}\operatorname{Tr}\Big(M\sum_{1\leq|i|\leq I} \rho_i e_i e_i^T\Big). \end{aligned}$$

On the other hand $\mathcal{L}u(x) = \langle p, \omega\rangle + \frac{1}{2}\operatorname{Tr}(MD)$. The announced result follows by identification since $p \in \mathbb{R}^d$ and $M \in S_d$ are arbitrary. □

Proposition 2.1.3 is stated in terms of the pair $(h\omega, D)$, because the multiplicative factor h arises naturally in the application to finite difference schemes; in contrast, we avoid this factor in Definition 2.1.2 for clarity. In dimension $d = 1$, and viewing 1-vectors and 1×1-matrices as scalars, one easily checks that (ω, D) is absolutely feasible iff $|\omega| \leq D$, and with the notations of Definition 2.1.2 one has $I = 1$, $e_1 = 1$, and $\rho_{\pm 1} = \frac{1}{2}(D \pm \omega)$, which corresponds to the usual centered finite differences scheme. Note that discretizing (2.1) using upwind finite differences for the first order term fails the consistency test on quadratic functions requested in Proposition 2.1.3. The following construction generalizes the centered finite differences scheme to dimension $d \in \{2, 3\}$.

Our numerical scheme relies on a tool from a lattice geometry known as *Selling's decomposition*, described in more detail in section 2.2.1, see also [Sel74; CS92]. It associates to each positive definite matrix $D \in S_d^{++}$, where $d \in \{2, 3\}$, a specific decomposition of the following form

$$D = \sum_{1\leq i\leq I} \sigma_i e_i e_i^\top, \qquad \text{where } \sigma_i \geq 0,\ e_i \in \mathbb{Z}^d,\ \forall 1 \leq i \leq I, \tag{2.4}$$

and $I := d(d + 1)/2$. Selling's decomposition has already been used in the design of difference schemes in dimension $d \in \{2, 3\}$, for (divergence form) anisotropic diffusion in [FM14], and for various anisotropic eikonal equations in [Mir18; Mir19]. It is at the foundation of degenerate elliptic and second order consistent discretizations of the fully non-linear two-dimensional Monge-Ampère [BCM16] and Pucci (Chapter 5) equations. In dimension $d = 2$, an equivalent construction based

on the Stern-Brocot dyadic tree of rational numbers is used in [BOZ04] for the Hamilton-Jacobi-Bellman equation of Stochastic control.

The support $(e_i)_{i=1}^I$ of Selling's decomposition, which is also the stencil of the numerical scheme proposed in this paper, tends to align with the anisotropy defined by the matrix D. This is illustrated on Figure 2.1, where we use the following parametrization (closely related with Pauli matrices in quantum mechanics) of the set of symmetric positive definite matrices of size two and with unit determinant:

$$D(a,b) := \frac{1}{\sqrt{1-a^2-b^2}} \begin{pmatrix} 1+a & b \\ b & 1-a \end{pmatrix}, \qquad a^2+b^2<1. \tag{2.5}$$

Definition 2.1.4 (Finite difference operators)**.** For any $e \in \mathbb{Z}^d$, $h>0$, $u : h\mathbb{Z}^d \to \mathbb{R}$, we let

$$\delta_h^e u(x) := \frac{u(x+he)-u(x-he)}{2h}, \qquad \Delta_h^e u(x) := \frac{u(x+he)-2u(x)+u(x-he)}{h^2}.$$

Given $D \in S_d^{++}$ where $d \in \{2,3\}$, with Selling decomposition $D = \sum_{1\le i\le I} \sigma_i e_i e_i^\top$, we let

$$\nabla_h^D u(x) = \sum_{1\le i\le I} \sigma_i\, \delta_h^{e_i} u(x)\, e_i, \qquad \Delta_h^D u(x) = \sum_{1\le i\le I} \sigma_i \Delta_h^{e_i} u(x).$$

The centered finite differences $\delta_h^e u(x) = \langle e, \nabla u(x)\rangle + \mathcal{O}(h^2)$ and second order finite differences $\Delta_h^e u(x) = \langle e, \nabla^2 u(x) e\rangle + \mathcal{O}(h^2)$, are classical constructs. In combination with Selling's decomposition, they are here used to define discrete anisotropic gradient and Laplacian operators, with the following consistency properties easily derived from (2.4)

$$\nabla_h^D u(x) = D\nabla u(x) + \mathcal{O}(h^2), \qquad \Delta_h^D u(x) = \mathrm{Tr}(D\nabla^2 u(x)) + \mathcal{O}(h^2). \tag{2.6}$$

For context, Selling's decomposition of the matrix $D = \mathrm{Id}$ yields up to permutation the canonical basis $(e_1, \cdots, e_d)$ with unit weights $\sigma_1 = 1, \cdots, \sigma_d = 1$, whereas the remaining weights vanish: $\sigma_i = 0$ for all $d < i \le I := d(d+1)/2$ (and the corresponding vectors e_i are not uniquely determined). As a result ∇_h^{Id} and Δ_h^{Id} are the classical finite differences discretizations of the gradient and Laplacian, whose stencil only involves the immediate grid neighbors.

Definition 2.1.5 (Canonical discretization)**.** We say that $(h\omega, D) \in \mathbb{R}^d \times S_d^{++}$, $d \in \{2,3\}$, is *canonically feasible* if the following operator L_h is DDE

$$-L_h u(x) := \langle D^{-1}\omega, \nabla_h^D u(x)\rangle + \tfrac{1}{2}\Delta_h^D u(x). \tag{2.7}$$

Equivalently, but more explicitly, (ω, D) is canonically feasible iff the following weights are non-negative

$$\rho_i := \frac{\sigma_i}{2}\big(1 + \langle \omega, D^{-1} e_i\rangle\big), \tag{2.8}$$

for all $1 \le |i| \le I$, where $D = \sum_{1\le i\le I} \sigma_i e_i e_i^T$ is Selling's decomposition (2.4) and thus $I := d(d+1)/2$. Note that these weights obey (2.3) by construction, which reflects the fact that for a quadratic function u the expansions (2.6) are exact (no remainder), so that (2.7) matches (2.1).

Definition 2.1.5 outlines a simple, canonical and practical discretization of the anisotropic linear PDE operator (2.1), often referred to as *our numerical scheme* in the paper. By construction, canonical feasibility implies absolute feasibility, but the latter can be achieved in a variety of other ways, using possibly a different number of terms I, a different support $(e_i)_{i=1}^I$, or different weights $(\rho_i)_{1\le|i|\le I}$. Note also that (2.7) is second order consistent with the PDE operator (2.1), in view of (2.6), whereas the conditions of Definition 2.1.2 only imply first order consistency. We next state the main result of this paper.

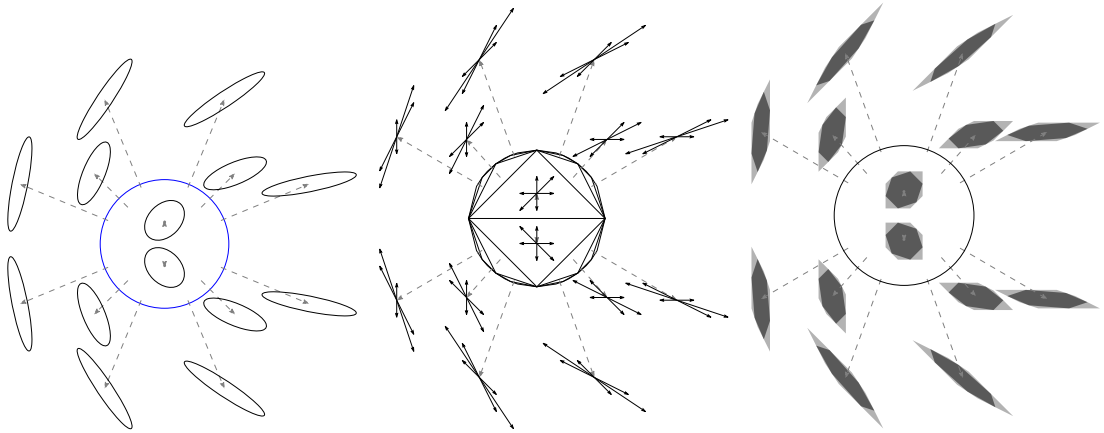


Figure 2.1: To each point (a, b) of the unit disk, we associate the matrix $D = D(a, b)$ defined by (2.5). *Left:* Ellipse defined by $\{\langle v, D^{-1}v\rangle \leq 1;\ v \in \mathbb{R}^2\}$. Points close to the unit disk boundary (shown blue) yield strongly anisotropic ellipses. *Center:* Support $(e_i)_{i=1}^I$ of Selling's decomposition, which is also the stencil of our finite difference scheme for the given anisotropy. *Right:* Set of vectors ω for which the pair (ω, D) is canonically feasible (dark gray), or absolutely feasible (dark and light gray), computed via Proposition 2.3.9. The scale of the three figures may not match.

Theorem 2.1.6. *Let $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$, where $d \in \{2, 3\}$. If (ω, D) is absolutely feasible, then $(c_d\,\omega, D)$ is canonically feasible, with $c_2 := 1/2$ and $c_3 := 1/6$.*

Taking the contraposition, Theorem 2.1.6 shows that if the canonical discretization of Definition 2.1.5 does not yield a DDE scheme in some practical instance, then (up to the factor c_d) obtaining a DDE scheme will require a serious compromise: either substantially reduce the grid scale (which amounts to multiplying ω by a small factor by homogeneity), or give up second order consistency (which allows using upwind finite differences, and eliminates the constraint of absolute feasibility). The following result in contrast provides a direct criterion for canonical feasibility. We denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the Euclidean norm and scalar product. Let also $\|e\|_M := \sqrt{\langle e, Me\rangle}$ and $\|A\| := \max_{\|x\|\leq 1} \|Ax\|$ for any $e \in \mathbb{R}^d$, $M \in S_d^{++}$, and matrix A.

Theorem 2.1.7. *Let $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$, where $d \in \{2, 3\}$, and let $M := D^{-1}$. If one has*

$$\|M\|^{\frac{1}{2}} \|\omega\|_M \leq c_d, \tag{2.9}$$

then (ω, D) is canonically feasible, with $c_2 := 1/2$ and $c_3 := 1/(4\sqrt{3})$.

The existence of a finite difference discretization, degenerate elliptic and second order consistent, is not the only practical concern: the width of the stencil used is also of importance. Excessively wide stencils reduce the effective discretization scale of the scheme, thus also the accuracy of the numerical results. They may also raise difficulties with the treatment of boundary conditions, computer parallelization, matrix conditioning and sparsity, etc. See also the discussion section 2.4. We provide two results related to the stencil width. First, we show that the canonical discretization has the smallest support of all possible DDE and second order consistent discretizations, in dimension two, in the strong sense of convex hull inclusion. We denote by $\mathrm{Hull}(E)$ the convex hull of a subset E of a vector space.

Theorem 2.1.8. *Let $(\omega, D) \in \mathbb{R}^2 \times S_2^{++}$ be canonically feasible, and let $(\rho_i, e_i)_{1\le|i|\le I}$ be the corresponding decomposition (2.8), pruned so that $\rho_i \neq 0$ or $\rho_{-i} \neq 0$ for all $1 \le i \le I$. Let $(\rho'_i, e'_i)_{1\le|i|\le I'}$ be another decomposition, as in Definition 2.1.2. Then*

$$\mathrm{Hull}\{e_i;\, 1 \le |i| \le I\} \subset \mathrm{Hull}\{e'_i;\, 1 \le |i| \le I'\}.$$

Second, we provide explicit bounds on the stencil width in terms of the differential operator coefficients and anisotropy.

Theorem 2.1.9. *Let $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$ be canonically feasible, where $d \in \{2, 3\}$, and let $(\rho_i, e_i)_{1\le|i|\le I}$ be the corresponding decomposition (2.8), where $I = d(d+1)/2$. Then $\|e_i\|_M \le C_d\sqrt{\|M\|}$ for all $1 \le i \le I$, where $M := D^{-1}$, and $C_2 = 2$ and $C_3 = 4\sqrt{3}$.*

Theorem 2.1.9 implies in particular that $\|e_i\| \le C_d\,\mathrm{Cond}(D)$, for all $1 \le i \le I$, where $\mathrm{Cond}(D) := \sqrt{\|D\|\|D^{-1}\|}$. See also [Mir16] for average case bounds in dimension $d = 2$, under random rotations $R_\theta^\top D R_\theta$ of the tensor, $\theta \in [0, 2\pi]$.

Outline

Section §2.2 is devoted to further discussion of the canonical discretization, and to the proofs of Theorems 2.1.7, 2.1.8 and 2.1.9 which follow rather directly from arguments presented in [Mir18] and [Mir16]. Section §2.3 establishes Theorem 2.1.6. Numerical experiments are presented in §2.4.

2.2 The canonical discretization

This section is devoted to a further presentation of the construction of Definition 2.1.5, here referred to as the *canonical discretization* of a second order linear PDE operator. We review Selling's algorithm in §2.2.1, finalizing the algorithmic description of our numerical scheme. We describe in §2.2.2 an interpretation of this algorithm as an optimization procedure, involving objects from the field of lattice geometry known as Voronoi's first reduction and Ryskov's polyhedron. Theorems 2.1.7, 2.1.8 and 2.1.9 are proved in §2.2.3.

The results presented §2.2.3 are new, whereas the more classical techniques described in §2.2.1 and §2.2.2 are required for completeness and as a preliminary to the proof of Theorem 2.1.6 in §2.3.

2.2.1 Selling's algorithm and formula

We describe Selling's algorithm [Sel74; CS92], and the related tensor decomposition formula which is invoked in Definition 2.1.5 of the numerical scheme considered in this paper.

Selling's algorithm

This algorithm belongs to the field of lattice geometry [NS04], which among other things studies coordinate systems in additive lattices (here $\mathbb{Z}^d$), adapted to the geometry defined by a given positive definite quadratic form (here defined by $D \in S_d^{++}$). The next definition introduces such a concept.

Definition 2.2.1. A superbase of $\mathbb{Z}^d$ is a $(d+1)$-tuple $b = (v_0, \cdots, v_d) \in (\mathbb{Z}^d)^{d+1}$ such that $|\det(v_1, \cdots, v_d)| = 1$ and $v_0 + \cdots + v_d = 0$. It is said D-obtuse, where $D \in S_d^{++}$, if $\langle v_i, Dv_j\rangle \le 0$ for all $0 \le i < j \le d$.

Given a positive definite tensor $D \in S_d^{++}$, where $d \in \{2,3\}$, Selling's algorithm constructs a D-obtuse superbase, see Algorithm 1. Note that the algorithm does not extend to dimension $d \geq 4$, and indeed there exists a matrix $D \in S_4^{++}$ for which no D-obtuse superbase exists [Sch09a].

Algorithm 1 Selling's algorithm

Input: A positive definite tensor $D \in S_d^{++}$, and a superbase $b = (v_0, \cdots, v_d)$, where $d \in \{2,3\}$.
While there exists $0 \leq i < j \leq d$ such that $\langle v_i, Dv_j \rangle > 0$ **do**

$$\text{If } d = 2,\ b \leftarrow (-v_i, v_j, v_i - v_j).$$
$$\text{If } d = 3,\ b \leftarrow (-v_i, v_j, v_i + v_k, v_i + v_l) \text{ where } \{k,l\} = \{0,1,2,3\} \setminus \{i,j\}.$$

Output: b, which is now a D-obtuse superbase.

Proof of correctness and termination of Algorithm 1. Denote by b the current superbase at the beginning of an iteration. If the stopping criterion holds, then b is D-obtuse, as desired. Otherwise, denoting by b' the updated superbase, one easily checks that

$$\mathcal{E}_D(b') = \mathcal{E}_D(b) - C_d \langle v_i, Dv_j \rangle \qquad \text{where} \quad \mathcal{E}_D(b) := \sum_{0 \leq k \leq d} \|v_k\|_D^2, \tag{2.10}$$

and where $C_2 = 4$ and $C_3 = 2$. Thus $\mathcal{E}_D(b') < \mathcal{E}_D(b)$. Since there exists only finitely many superbases b such that $\mathcal{E}_D(b)$ is below a given constant, Selling's algorithm must terminate. □

Selling's algorithm is not the only means to produce a D-obtuse superbase. For instance Corollary 1 and Proposition 1 in [FM14] show in dimension $d \in \{2,3\}$ how to produce a D-obtuse superbase from another type of system of coordinates referred to as D-reduced basis, resulting in a $\mathcal{O}(\ln(\|D\|\|D^{-1}\|))$ numerical complexity [NS04]. Selling's algorithm is however efficient enough for applications to PDE discretization, which usually involve moderate condition numbers, and therefore it is used in all our numerical experiments §2.4.

Selling's decomposition

This mathematical formula allows, once a D-obtuse superbase of $\mathbb{Z}^d$ is known, to decompose the tensor $D \in S_d^{++}$ in the form of (2.4). For that purpose, we associate to each superbase a family of vectors $(e_{ij})_{i \neq j}$ defined by duality relations.

Definition 2.2.2. Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then for any i, j in $\{0, \cdots, d\}$ such that $i \neq j$ we let $e_{ij} \in \mathbb{Z}^d$ be the unique vector obeying $\langle e_{ij}, v_k \rangle := \delta_{ik} - \delta_{jk}$, for all $0 \leq k \leq d$.

Note that Definition 2.2.2 characterizes $e_{ij} \in \mathbb{R}^d$ by $d+1$ linear relations. This does make sense in view of the redundancy $v_0 + \cdots + v_d = 0$ of the linear forms, and of the compatibility $(\delta_{i0} - \delta_{j0}) + \cdots + (\delta_{id} - \delta_{jd}) = 1 - 1 = 0$ of the right-hand sides. The vectors e_{ij} admit explicit expressions when $d \in \{2,3\}$, namely (up to the sign)

$$e_{ij} = \pm v_k^{\perp} \ \text{ if } d = 2, \qquad (\text{resp. } e_{ij} = \pm v_k \times v_l \ \text{ if } d = 3), \tag{2.11}$$

where $\{i,j,k\} = \{0,1,2\}$ (resp. $\{i,j,k,l\} = \{0,1,2,3\}$). For all $i, j, k, l \in \{0, \cdots, d\}$ such that $i \neq j$ and $k \neq l$ one also has the useful identity

$$\big\langle v_k, (e_{ij} e_{ij}^{\top}) v_l \big\rangle = \big\langle e_{ij}, (v_k \otimes v_l) e_{ij} \big\rangle = \langle e_{ij}, v_k \rangle \langle e_{ij}, v_l \rangle \tag{2.12}$$

$$= \begin{cases} -1 \text{ if } \{i,j\} = \{k,l\}, \\ 0 \text{ otherwise.} \end{cases}$$

We denoted by $v \otimes w := \frac{1}{2}(vw^\top + wv^\top) \in S_d$ the symmetrized outer product of two vectors $v, w \in \mathbb{R}^d$. The next lemma shows how a superbase of $\mathbb{Z}^d$ defines a decomposition of an arbitrary tensor D, involving integer offsets. If the superbase is D-obtuse, then the weights are non-negative, and the decomposition is known as Selling's decomposition or *formula* [Sel74; CS92].

Lemma 2.2.3 (Selling's decomposition)**.** *Let $D \in S_d$, and let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then*

$$D = -\sum_{0 \le i < j \le d} \langle v_i, Dv_j \rangle e_{ij} e_{ij}^\top. \tag{2.13}$$

If $D \in S_d^{++}$ and b is D-obtuse, then (2.13) *is known as* Selling's decomposition *of D.*

Proof. Denote by D' the r.h.s. of (2.13). By (2.12) we obtain $\langle v_k, Dv_l \rangle = \langle v_k, D'v_l \rangle$ for all $0 \le k < l \le d$. These $d(d+1)/2$ independent linear relations imply $D = D'$, as announced. □

We finally complete the description of our numerical scheme construction, see Definition 2.1.5. Given a positive definite tensor D, build a D-obtuse superbase using Selling's algorithm or another method. Then Selling's formula (2.13) yields the required tensor decomposition $D = \sum_{1 \le i \le I} \sigma_i e_i e_i^\top$ with $I = d(d+1)/2$, $\sigma_i \ge 0$, $e_i \in \mathbb{Z}^d \setminus \{0\}$. We emphasize that one cannot replace Selling's formula with another tensor decomposition in Definition 2.1.5, or Theorems 2.1.6, 2.1.7, 2.1.8 and 2.1.9 would fail. Finally, let us mention that Selling's decomposition is uniquely determined by the tensor D, and thus independent of the choice of D-obtuse superbase, see Remark 2.2.13.

2.2.2 Ryskov's polyhedron and Voronoi's first reduction

We introduce two concepts from lattice geometry, Ryskov's polyhedron and Voronoi's first reduction [Sch09a], allowing us to rephrase Selling's algorithm as a simplex-like optimization method solving a linear program. In order to prevent any confusion, let us insist that these geometric tools are *not* connected with the classical concept of Voronoi diagram, which is instead related with Voronoi's *second* reduction [Sch09a]. Ryskov's polyhedron is an unbounded subset $\mathcal{M}_d \subset S_d$, defined as follows[1]

$$\mathcal{M}_d := \{M \in S_d;\ \forall e \in \mathbb{Z}^d \setminus \{0\},\ \langle e, Me \rangle \ge 1\}. \tag{2.14}$$

Remark 2.2.4 (Identification of duplicate constraints). The constraints associated in (2.14) with a vector $e \in \mathbb{Z}^d \setminus \{0\}$ and with its opposite $-e$ are obviously equivalent. We regard them as a single constraint, associated with the equivalence class $\pm e$.

The main result proved in this subsection is the classification of the edges and vertices of the polyhedron $\mathcal{M}_d$ in dimension $d \in \{2, 3\}$, see Corollary 2.2.11. These objects are actually known in all dimensions $d \le 8$ see [Sch09b; CS88; DSV07], hence the results presented in this subsection are not new. The proof is recalled for completeness and because its arguments are adapted in §2.3.1 for the proof of Theorem 2.1.6.

[1]Depending on the author, Ryskov's polyhedron (2.14) is defined via the constraints $\langle e, Me \rangle \ge \lambda$, $e \in \mathbb{Z}^d \setminus \{0\}$, where λ is one, two, or an unspecified positive constant [Sch09b]. These definitions are equivalent up to a homothety of $\mathcal{M}_d$.

Regularity of Ryskov's polyhedron

We refer to Appendix 2.B for some general terminology on polyhedra and linear programming. Recall that, by Minkowski's convex body theorem [Sch09a], any centrally symmetric convex body $K \subset \mathbb{R}^d$ of volume $\mathrm{Vol}(K) > 2^d$ contains a point of $\mathbb{Z}^d \setminus \{0\}$.

Lemma 2.2.5. *Each $M \in \mathcal{M}_d$ is positive definite, and $\det(M) \geq c_d$ where $c_d > 0$ is a constant.*

Proof. If $M \in \mathcal{M}_d$, then by construction M is positive semi-definite and the set $K = \{x \in \mathbb{R}^d; \langle x, Mx\rangle < 1\}$ contains no point of $\mathbb{Z}^d \setminus \{0\}$. By Minkowski's convex body theorem one has $2^d \geq \mathrm{Vol}(K) = \mathrm{Vol}(B)\det(M)^{-\frac{1}{2}}$, where B denotes the Euclidean unit ball, as announced. The announced result thus holds with (sub-optimal) constant $c_d := \mathrm{Vol}(B)^2/2^{2d}$. □

The optimal constant in Lemma 2.2.5 is $c_d = \gamma_d^{-d}$, where γ_d is known as Hermite's constant [Sch09a].

Corollary 2.2.6. *The polyhedron $\mathcal{M}_d$ is regular in the sense of Definition 2.B.1.*

Proof. Let us check the three points of this definition. (i) The set $\mathcal{M}_d$ contains all $M \in S_d$ such that $M \succeq \mathrm{Id}$, hence it has non-empty interior, as required. (ii) The defining constraints obey $\mathrm{Span}\{ee^\top; e \in \mathbb{Z}^d \setminus \{0\}\} = \mathbb{S}_d$, as required. (iii) For any $M, M' \in S_d$ and any $e \in \mathbb{R}^d$ one has $\langle e, M'e\rangle \geq (\lambda_{\min}(M) - \|M - M'\|)\|e\|^2$, where $\lambda_{\min}(M) > 0$ denotes the smallest eigenvalue. Given $M \in S_d^{++}$, one thus has $\langle e, M'e\rangle > 1$ whenever $\|M' - M\| < \lambda_{\min}(M)/2$ and $\|e\| \geq 2/\lambda_{\min}(M)$. This shows that only finitely many constraints defining the polyhedron $\mathcal{M}_d$ active in the neighborhood of any $M \in \mathcal{M}_d$, as required. □

Vertices and edges of Ryskov's polyhedron

We describe a family of vertices of $\mathcal{M}_d$ in Lemma 2.2.8, the corresponding edges in Lemma 2.2.10, $d \in \{2, 3\}$, and show in Corollary 2.2.11 that this exhausts the skeleton of $\mathcal{M}_d$.

Definition 2.2.7. To each superbase $b = (v_0, \cdots, v_d)$ of $\mathbb{Z}^d$ one associates the matrix

$$M_b = \frac{1}{2}\sum_{0\leq i\leq d} v_i v_i^\top. \tag{2.15}$$

Lemma 2.2.8. *Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then $\langle e, M_b e\rangle \geq 1$ for all $e \in \mathbb{Z}^d \setminus \{0\}$, with equality iff $e = e_{ij}$ for some $i, j \in \{0, \cdots, d\}$, $i \neq j$, see Definition 2.2.2.*

Proof. Let $e \in \mathbb{Z}^d \setminus \{0\}$ and $S := 2\langle e, M_b e\rangle = \sum_{0\leq i\leq d}\langle v_i, e\rangle^2$. Then S is the sum of the squares of the integers $\langle v_i, e\rangle$, $0 \leq i \leq d$, which are not all zero, and obey $\sum_{0\leq i\leq d}\langle v_i, e\rangle = \langle 0, e\rangle = 0$. Thus $S \geq 2$, with equality iff there exists $i \neq j$ such that $\langle v_i, e\rangle = 1$, $\langle v_j, e\rangle = -1$, and $\langle v_k, e\rangle = 0$ for all $k \notin \{i, j\}$. In other words $e = e_{ij}$, as announced. □

By Lemma 2.2.8, one has $M_b \in \mathcal{M}_d$ for any superbase b. Furthermore, M_b saturates the $d(d+1)/2 = \dim(S_d)$ linearly independent constraints associated with the vectors $\pm e_{ij}$, where $0 \leq i < j \leq d$, and satisfies strictly the constraints associated with any other $e \in \mathbb{Z}^d \setminus \{0\}$. This shows that M_b is a non-degenerate vertex of the polyhedron $\mathcal{M}_d$. The edges emanating from this vertex, in dimension $d \in \{2, 3\}$, are described in Lemma 2.2.10 below.

We introduce in the next definition an adjacency relation on the set of superbases of $\mathbb{Z}^d$, which is reminiscent of the superbase updates involved in Selling's algorithm, Algorithm 1. This similarity is not by accident, and it leads to an interpretation of Selling's algorithm as a linear program solver, see Proposition 2.2.12.

Definition 2.2.9. One defines the following adjacency relations for superbases of $\mathbb{Z}^d$, $d \in \{2,3\}$,

$$\begin{aligned}(v_0,v_1,v_2) &\leftrightarrow (-v_0,v_1,v_0-v_1),\\ (v_0,v_1,v_2,v_3) &\leftrightarrow (-v_0,v_1,v_2+v_0,v_3+v_0),\end{aligned} \tag{2.16}$$

and likewise up to a permutation and/or a global change of sign of the superbase.

Lemma 2.2.10. *Let $d \in \{2,3\}$ and let b be a superbase of $\mathbb{Z}^d$. The edges of $\mathcal{M}_d$ containing M_b coincide with the segments $[M_b, M_{b'}]$, where b' is a superbase of $\mathbb{Z}^d$ adjacent to b.*

Proof. Recall that M_b is a non-degenerate vertex of $\mathcal{M}_d$. Therefore there exists $d(d+1)/2 = \dim(S_d)$ edges of $\mathcal{M}_d$ containing M_b, which are obtained by relaxing one of the constraints active at M_b (see also section 2.B.3 on this topic). In other words, the edges of $\mathcal{M}_d$ containing M_b can be parametrized by $0 \le \alpha < \beta \le d$ and obtained as

$$E_{\alpha\beta} = \{M \in \mathcal{M}_d;\ \langle e_{ij}, Me_{ij}\rangle = 1,\ 0 \le i < j \le d,\ (i,j) \neq (\alpha,\beta)\}.$$

Let b, b' be superbases of $\mathbb{Z}^d$ as in (2.16). Then distinguishing dimensions we compute

$$(d=2): \quad 2(M_{b'} - M_b) = (v_0-v_1)(v_0-v_1)^\top - v_2 v_2^\top = -4 v_0 \otimes v_1. \tag{2.17}$$

$$\begin{aligned}(d=3): \quad & 2(M_{b'} - M_b) \\ &= (v_2+v_0)(v_2+v_0)^\top + (v_3+v_0)(v_3+v_0)^\top - v_2 v_2^\top - v_3 v_3^\top \\ &= 2 v_0 \otimes (v_0+v_2+v_3) = -2 v_0 \otimes v_1.\end{aligned} \tag{2.18}$$

The symmetrized outer product $\otimes$ was introduced in (2.12). Thus elements M in $[M_b, M_{b'}]$ obey the constraints $\langle e_{ij}, Me_{ij}\rangle = 1$ whenever $\{i,j\} \neq \{0,1\}$, by (2.12). Therefore $[M_b, M_{b'}] \subset E_{01}$, and equality holds since M_b and $M_{b'}$ are vertices of $\mathcal{M}_d$ and $E_{01} \subset \partial\mathcal{M}_d$. Likewise, by permuting the indices, we obtain for all $0 \le i < j \le d$ an edge of $\mathcal{M}_d$ of the form $[M_b, M_{b'}]$ where b' is adjacent to b, obeying all the constraints active at the non-degenerate vertex M_b but the one associated with $\pm e_{ij}$ (previously $\pm e_{01}$). □

Corollary 2.2.11. *The vertices (resp. bounded edges) of Ryskov's polyhedron $\mathcal{M}_d$, where $d \in \{2,3\}$, take the form M_b where b is a superbase of $\mathbb{Z}^d$ (resp. $[M_b, M_{b'}]$ where b' is an adjacent superbase). There are no unbounded edges in $\partial\mathcal{M}_d$.*

Proof. The result follows from Lemma 2.2.10, and the fact that the graph defined by the vertices and edges of a regular polyhedron is connected. □

Voronoi's first reduction

Voronoi's first reduction $\mathrm{Vor}(D)$, of a positive definite quadratic form $D \in S_d^{++}$, is defined as a linear minimization problem over Ryskov's polyhedron

$$\mathrm{Vor}(D) := \inf_{M \in \mathcal{M}_d} \mathrm{Tr}(DM). \tag{2.19}$$

This linear program, in dimension $d(d+1)/2$ and subject to infinitely many constraints, is well posed as shown by Voronoi himself [Vor08; Sch09a], in the sense that the collection of minimizers is non-empty and compact (generically it is a point) for any $D \in S_d^{++}$. The next proposition reproves this fact in dimension $d \in \{2,3\}$.

Proposition 2.2.12. *Let $D \in S_d^{++}$, where $d \in \{2,3\}$. Then Voronoi's first reduction is a well posed linear program, attaining its minimum at vertices M_b of $\mathcal{M}_d$ associated with a D-obtuse superbase b.*

Proof. By lemma 2.2.10, Selling's algorithm defines a walk on the graph defined by the vertices and edges of Ryskov's polyhedron. Observing that $\mathcal{E}_D(b) = 2\,\mathrm{Tr}(DM_b)$, see (2.10) and (2.15), we see that the next vertex selection reduces the linear program's objective function, whenever that is possible. Compare also (2.10, left) with (2.17) and (2.18). Since Selling's algorithm terminates, it solves the linear program (2.19), by the general results in §2.B.2. Furthermore, by Definition 2.2.1, it terminates precisely when reaching a D-obtuse superbase, which concludes the proof. □

Note that the proof of the previous proposition outlines a close relationship between Selling's algorithm and the simplex algorithm [BG15] applied to the linear program (2.19).

Remark 2.2.13 (Uniqueness of Selling's decomposition). Consider the decomposition (2.13) of a tensor $D \in S_d^{++}$, associated with a D-obtuse superbase b (if any exists, which is only guaranteed in dimension $d \leq 3$). By Lemma 2.2.8, it can be rephrased as a set of KKT relations for the linear program (2.19) at $M_b \in \mathcal{M}_d$, see Definition 2.B.5. Since M_b is a non-degenerate vertex of $\mathcal{M}_d$, the coefficients of this KKT relation are uniquely determined, even if there is no uniqueness of the D-obtuse superbase, see Proposition 2.B.6. In contrast, Voronoi's reduction (2.19) in dimension $d \geq 4$, or our variant $\widetilde{\mathrm{Vor}}(\omega, D)$ introduced §2.3.2 in dimension $d \geq 2$, involve polyhedra with degenerate vertices, at which the KKT relations are often non-uniquely determined.

2.2.3 Proof of Theorems 2.1.7, 2.1.8 and 2.1.9

Theorems 2.1.7 and 2.1.9, announced in the introduction, provide respectively a criterion for the existence of our discretization, and an estimate of the size of its support. They both follow from the next lemma, which bounds the norm of the vectors defined dually from an obtuse superbase.

Lemma 2.2.14 (Corollary 4.12 in [Mir18])**.** *Let $D \in S_d^{++}$ where $d \in \{2, 3\}$. Let b be a D-obtuse superbase, and let $e = e_{ij}$, for some $i, j \in \{0, \cdots, d\}$ such that $i \neq j$, see Definition 2.2.2. Then, denoting $C_2 := 2$ and $C_3 := 4\sqrt{3}$, one has*

$$\|e\|_M \leq C_d \|M\|^{\frac{1}{2}}, \qquad \text{where} \quad M := D^{-1}. \tag{2.20}$$

We refer to [Mir18] for the proof of Lemma 2.2.14, and use this result here to establish Theorems 2.1.7 and 2.1.9.

Proof of Theorem 2.1.9. Recall that the numerical scheme construction in Definition 2.1.5 relies on Selling's decomposition of a tensor $D \in S_d^{++}$, see Lemma 2.2.3. In particular the offsets $(e_i)_{1 \leq i \leq I}$ with $I = d(d+1)/2$ appearing in Theorem 2.1.9, are (up to reindexing) the same as those appearing in Lemma 2.2.14 and there denoted $(e_{ij})_{0 \leq i < j \leq d}$. The announced result follows from (2.20). □

Proof of Theorem 2.1.7. Denote $M := D^{-1}$ and let $e = e_{ij}$ for some $0 \leq i < j \leq d$, with the notations of Lemma 2.2.14. Then

$$|\langle \omega, D^{-1} e \rangle| = |\langle \omega, Me \rangle| \leq \|\omega\|_M \|e\|_M \leq C_d \|M\|^{\frac{1}{2}} \|\omega\|_M.$$

Condition (2.9) thus implies that $|\langle \omega, D^{-1} e \rangle| \leq 1$, and therefore that the weights (2.8) are non-negative, which as announced proves the absolute feasibility of (ω, D). □

The rest of this section is devoted to the proof of Theorem 2.1.8. For that purpose, we need to introduce the geometrical concept of Voronoi vector [Sch09a].

Definition 2.2.15. A point $e \in \mathbb{Z}^d \setminus \{0\}$ is an M-Voronoi vector, where $M \in S_d^{++}$, if there exists $p \in \mathbb{R}^d$ (referred to as the witness) such that

$$\|p - 0\|_M = \|p - e\|_M \leq \|p - x\|_M, \text{ for all } x \in \mathbb{Z}^d. \tag{2.21}$$

One says that e is a *strict* M-Voronoi vector if the above inequality is strict for all $x \notin \{0, e\}$.

The origin 0 is introduced in (2.21, left) to emphasize the geometrical interpretation. In the language of Voronoi diagrams, e is an M-Voronoi vector iff the Voronoi cells of 0 and e intersect, in the diagram of $\mathbb{R}^d$ associated with the sites $\mathbb{Z}^d$ and metric $\|\cdot\|_M$. The (strict) M-Voronoi vectors can be determined from an M-obtuse superbase, as shown by the next lemma in dimension $d = 2$. See Theorem 3 in [CS92] for a related argument in arbitrary dimension.

Lemma 2.2.16. *Let $M \in S_2^{++}$ and let e_0, e_1, e_2 be an M-obtuse superbase. Then $\pm e_0, \pm e_1, \pm e_2$ are M-Voronoi vectors. Furthermore e_0 is a strict M-Voronoi vector iff $\langle e_1, Me_2\rangle < 0$ (likewise for $-e_0$, and likewise permuting (e_0, e_1, e_2)).*

Proof. We first show, w.l.o.g., that e_0 is an M-Voronoi vector, whose witness is $p := e_0/2$. Note that $\|p - 0\|_M = \|p - e_0\|_M (= \|e_0/2\|_M)$ as required (2.21). Let $x \in \mathbb{Z}^2$ be arbitrary. Since $\det(e_1, e_2) = 1$, there exists $a, b \in \mathbb{Z}$ such that $x = ae_1 + be_2$. From this point a direct computation yields (2.21), as announced

$$\begin{aligned}
\|p - x\|_M^2 - \|p\|_M^2 &= \|(a + 1/2)e_1 + (b + 1/2)e_2\|_M^2 - \|(e_1 + e_2)/2\|_M^2 \\
&= (a^2 + a)\|e_1\|_M^2 + (b^2 + b)\|e_2\|_M^2 + (2ab + a + b)\langle e_1, Me_2\rangle \\
&\geq \Big((a^2 + a) + (b^2 + b) - (2ab + a + b)\Big)(-\langle e_1, Me_2\rangle) \\
&= -(a - b)^2\langle e_1, Me_2\rangle \ \geq 0.
\end{aligned}$$

In the third line we used $\|e_1\|_M^2 = \langle -e_0 - e_2, Me_1\rangle \geq -\langle e_1 Me_2\rangle$, and likewise $\|e_2\|_M^2 \geq -\langle e_1 Me_2\rangle$. In the rest of the proof, we show that e_0 is a strict M-Voronoi vector, under the additional assumption that $\langle e_1, Me_2\rangle < 0$. Indeed, if $\|p\| = \|p - x\|$, then $a = b$ by the above, thus $x = -ae_0$ and therefore $\|e_0/2\|_M = \|(a + 1/2)e_0\|_M$. This implies $a \in \{0, -1\}$, hence $x \in \{0, e_0\}$, and therefore e_0 is a strict M-Voronoi vector, as announced. □

Lemma 2.2.17. *Let $D \in S_2^{++}$ and let (e_0, e_1, e_2) be a D-obtuse superbase. Let $M := D^{-1}$ and $(v_0, v_1, v_2) = (e_0^\perp, e_1^\perp, e_2^\perp)$. Then (v_0, v_1, v_2) is an M-obtuse superbase. In addition, for any $i \neq j$ one has $\langle e_i, De_j\rangle < 0$ iff $\langle v_i, Mv_j\rangle < 0$.*

Proof. By construction one has $v_0, v_1, v_2 \in \mathbb{Z}^2$, $v_0 + v_1 + v_2 = (e_0 + e_1 + e_2)^\perp = 0$, and $\det(v_1, v_2) = \det(e_1, e_2) = \pm 1$. Thus (v_0, v_1, v_2) is a superbase of $\mathbb{Z}^2$. On the other hand, the obtuseness properties come from the following identity: for any $e, e' \in \mathbb{R}^2$, $D \in S_2^{++}$ and $M := D^{-1}$ one has

$$\langle e^\perp, Me'^\perp\rangle = \det(M)\langle e, De'\rangle.$$

(In the special case $D = \mathrm{Id}$, this identity expresses that rotation by $\pi/2$ is an isometry. In the general case $D = A^T A$ for some $A \in \mathrm{GL}_2(\mathbb{R})$, it follows from a linear change of variables and the relation $(Ae)^\perp = \mathrm{cof}(A)e^\perp$ where $\mathrm{cof}(A)$ denotes the cofactor matrix.) □

We are ready to prove Theorem 2.1.8, by adapting a result of [Mir16], devoted to operators without a first order term, and stated in terms of Voronoi vectors.

Lemma 2.2.18 (Adapted from Theorem 1.3 in [Mir16]). *Let $D \in S_2^{++}$, and let*

$$D = \sum_{1 \leq i \leq I} \sigma_i e_i e_i^\top$$

be the decomposition associated with a D-obtuse superbase by Lemma 2.2.3, pruned so that $\sigma_i \neq 0$ for all $1 \leq i \leq I$. Let also $D = \sum_{1 \leq i \leq I'} \sigma'_i e'_i e'^\top_i$ be another decomposition, with $I' > 0$, $\sigma'_i \geq 0$, $e'_i \in \mathbb{Z}^2 \setminus \{0\}$ for all $1 \leq i \leq I'$. Then

$$\mathrm{Hull}\{\pm e_i; 1 \leq i \leq I\} \subset \mathrm{Hull}\{\pm e'_i; 1 \leq i \leq I'\}.$$

Proof. Theorem 1.3 in [Mir16] provides a similar statement, except that the vectors $(\pm e_i)_{1 \leq i \leq I}$ are defined as the strict M-Voronoi vectors, where $M = D^{-1}$. By Lemmas 2.2.16 and 2.2.17, the tensor decomposition here considered (2.13) is also supported on the set of strict M-Voronoi vectors, and the result follows. □

Proof of Theorem 2.1.8. We use the notations of Theorem 2.1.8, and define $\sigma_i := \rho_i + \rho_{-i}$ for all $1 \leq i \leq I$, and $\sigma'_i := \rho_i + \rho'_{-i}$ for all $1 \leq i \leq I'$. Note that $\sigma_i > 0$ for all $1 \leq i \leq I$, since $\rho_i \neq 0$ or $\rho_{-i} \neq 0$ and both are non-negative. Then $D = \sum_{1 \leq i \leq I} \sigma_i e_i e_i^\top = \sum_{1 \leq i \leq I} \sigma'_i e'_i e'^\top_i$, and by Definition 2.1.5 the first decomposition comes from a D-obtuse superbase as in Lemma 2.2.3. Applying Lemma 2.2.18, and recalling that $e_{-i} := -e_i$, we conclude the proof of Theorem 2.1.8. □

2.3 Proof of Theorem 2.1.6

We establish in this section our main result, Theorem 2.1.6, on a compatibility relation needed for constructing our numerical scheme. This obstruction relates the grid scale h (safely ignored in this section), with the first order term $\omega \in \mathbb{R}^d$ and the second order term $D \in S_d^{++}$ of the discretized linear differential operator. More precisely, this result states that if (ω, D) is absolutely feasible (some discretization exists), then $(c_d\omega, D)$ is canonically feasible (our discretization exists), where $d \in \{2, 3\}$ and $c_d \in]0, 1]$ is a constant.

The guiding principle of the proof is to adapt to the pair $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$, of a vector and a symmetric positive definite matrix, the tools and techniques presented in §2.2.1 and §2.2.2, which originally apply to a matrix $D \in S_d^{++}$ alone. The arguments are split into three parts, and proceed as follows. We define and describe in section 2.3.1 a variant $\widetilde{\mathcal{M}}_d \subset \mathbb{R}^d \times S_d$ of Ryskov's polyhedron $\mathcal{M}_d \subset S_d$, see (2.14), involving an asymmetric perturbation of the constraints. The corresponding generalization $\widetilde{\mathrm{Vor}}(\omega, D)$ of Voronoi's first reduction $\widetilde{\mathrm{Vor}}(D)$, see (2.19), is discussed in §2.3.2. We conclude the proof of Theorem 2.1.6 in section 2.3.3, by studying a low dimensional linear feasibility problem.

2.3.1 A variant of Ryskov's polyhedron

We study of a variant of Ryskov's polyhedron (2.14). Denoted $\widetilde{\mathcal{M}}_d \subset \mathbb{R}^d \times S_d$, it is defined as follows

$$\widetilde{\mathcal{M}}_d := \{(\eta, M) \in \mathbb{R}^d \times S_d;\ \forall e \in \mathbb{Z}^d \setminus \{0\}, \langle \eta, e \rangle + \langle e, Me \rangle \geq 1\}. \tag{2.22}$$

This subsection is devoted to description of the vertices and edges of $\widetilde{\mathcal{M}}_d$, when $d \in \{2, 3\}$, see Theorem 2.3.1 below (no other result from this section is used in the following ones). Surprisingly enough, this structure is only barely richer than that of Ryskov's original polyhedron, see Corollary 2.2.11, despite the higher dimension.

The concepts of superbase b of $\mathbb{Z}^d$, the associated matrix $M_b \in S_d^{++}$, and the notion of adjacent superbases (b, b'), were introduced in Definitions 2.2.1, 2.2.7, and 2.2.9 respectively. Regular polyhedra and their edges are introduced in Definitions 2.B.1 and 2.B.2 of Appendix 2.B.

Theorem 2.3.1. *Let $d \in \{2, 3\}$. Then $\widetilde{\mathcal{M}}_d$ is a regular polyhedron, with:*

(a) Vertices: $(0, M_b)$, for all superbases b of $\mathbb{Z}^d$.

(b) Bounded edges: $[(0, M_b), (0, M_{b'})]$, for all adjacent superbases b and b' of $\mathbb{Z}^d$.

(c) Unbounded edges: $\{(0, M_b) + \lambda(v_I, v_I v_I^\top);\ \lambda \geq 0\}$, for all superbases b of $\mathbb{Z}^d$ and all $I \subsetneq \{0, \cdots, d\}$, $I \neq \emptyset$, where $b = (v_0, \cdots, v_d)$ and $v_I := \sum_{i \in I} v_i$.

The rest of this section is devoted to the proof of Theorem 2.3.1, following a line of arguments similar to the proof of Corollary 2.2.11. For commodity, we introduce a scalar product on $\mathbb{R}^d \times S_d$, as well as a family of elements $l_e \in \mathbb{R}^d \times S_d$, $e \in \mathbb{Z}^d \setminus \{0\}$, defined as follows:

$$\langle\!\langle (\eta, M), (\omega, D) \rangle\!\rangle := \langle \eta, \omega \rangle + \mathrm{Tr}(MD), \qquad l_e := (e, ee^\top). \tag{2.23}$$

By construction $\langle\!\langle l_e, (\eta, M) \rangle\!\rangle = \langle e, \eta \rangle + \langle e, Me \rangle$, which is convenient in view of (2.22). Observe that for any $\lambda_1, \cdots, \lambda_I, \mu_1, \cdots, \mu_I \in \mathbb{R}$ and $e_1, \cdots e_I \in \mathbb{Z}^d$, one has

$$\sum_{1 \leq i \leq I} \frac{\lambda_i + \mu_i}{2} \Big(e_i, e_i e_i^\top \Big) + \frac{\lambda_i - \mu_i}{2} \Big(-e_i, (-e_i)(-e_i)^\top \Big) = \left(\sum_{1 \leq i \leq I} \mu_i e_i, \sum_{1 \leq i \leq I} \lambda_i e_i e_i^\top \right). \tag{2.24}$$

Remark 2.3.2 (Erdahl's cone of quadratic functions). The set (2.22) is reminiscent of Erdahl's cone [Erd92; DSV12], another inhomogeneous generalization of Voronoi's constructions, defined as follows:

$$\mathcal{E}_d := \{f \text{ quadratic function on } \mathbb{R}^d;\ \forall e \in \mathbb{Z}^d, f(e) \geq 0\}$$

Recall that a quadratic function on is a map of the form $x \in \mathbb{R}^d \mapsto \alpha + \langle \eta, x \rangle + \langle x, Mx \rangle$. Thus for any $f \in \mathcal{E}_d$, the normalized function $f/f(0)$ (assuming $f(0) \neq 0$) can be identified with an element of

$$\{(\eta, M) \in \mathbb{R}^d \times S_d;\ \forall e \in \mathbb{Z}^d \setminus \{0\},\ \langle \eta, e \rangle + \langle e, Me \rangle \geq -1\}. \tag{2.25}$$

Despite the apparent similarity between (2.25) and (2.22), the set $\widetilde{\mathcal{M}}_d$ only resembles Erdahl's cone superficially. The set $\widetilde{\mathcal{M}}_d$ is more closely related with Ryskov's original polyhedron $\mathcal{M}_d$, as shown by Theorem 2.3.1 and Corollary 2.2.11.

Lemma 2.3.3. *For all $(\eta, M) \in \widetilde{\mathcal{M}}_d$ one has $M \in \mathcal{M}_d$.*

Proof. One has $\langle e, Me \rangle = \frac{1}{2}(\langle \eta, e \rangle + \langle e, Me \rangle) + \frac{1}{2}(\langle \eta, -e \rangle + \langle -e, M(-e) \rangle) \geq 1$, $\forall e \in \mathbb{Z}^d \setminus \{0\}$. □

Lemma 2.3.4. *The polyhedron $\widetilde{\mathcal{M}}_d$ is regular, in the sense of Definition 2.B.1.*

Proof. (i) Let $(\eta, M) \in \mathbb{R}^d \times S_d$ be such that $\|\eta\| \leq 1$ and $M \succeq 2\,\mathrm{Id}$. Then for any $e \in \mathbb{Z}^d \setminus \{0\}$ one has $\langle \eta, e \rangle + \langle e, Me \rangle \geq -\|e\| + 2\|e\|^2 \geq 1$ since $\|e\| \geq 1$. Thus $(\eta, M) \in \widetilde{\mathcal{M}}_d$, and therefore $\widetilde{\mathcal{M}}_d$ has a non-empty interior. (ii) Recalling that $\mathrm{Span}\{ee^\top; e \in \mathbb{Z}^d \setminus \{0\}\} = S_d$, see Lemma 2.2.3, and using (2.24) one obtains $\mathrm{Span}\{(e, ee^\top); e \in \mathbb{Z}^d \setminus \{0\}\} = \mathbb{R}^d \times S_d$, as required. (iii) Let $(\eta, M) \in \widetilde{\mathcal{M}}_d$. Then $M \in \mathcal{M}_d$, by Lemma 2.3.3, and therefore M is a symmetric positive definite matrix, whose smallest eigenvalue is here denoted $\lambda_{\min}(M) > 0$. Then for any (η', M') such that $\|\eta - \eta'\| \leq 1$ and $\|M - M'\| \leq \lambda_{\min}(M)/2$ one has for all $e \in \mathbb{Z}^d \setminus \{0\}$

$$\langle \eta', e \rangle + \langle e, M'e \rangle \geq -\|\eta'\|\|e\| + (\lambda_{\min}(M) - \|M - M'\|)\|e\|^2$$

$$\geq (\lambda_{\min}(M)\|e\|/2 - \|\eta\| - 1)\|e\|.,$$

It follows that $\langle \eta', e\rangle + \langle e, M'e\rangle \geq 2$ if $\|e\| \geq 2(\|\eta\| + 3)/\lambda_{\min}(M)$. This shows that only finitely many of the constraints defining the polyhedron $\widetilde{\mathcal{M}}_d$ are active in the neighborhood of $(\eta, M) \in \widetilde{\mathcal{M}}_d$, as required. □

The next lemma describes a family of vertices of $\widetilde{\mathcal{M}}_d$.

Lemma 2.3.5. *For any vertex M of $\mathcal{M}_d$, the pair $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$. In addition, the active constraints at a vertex $M \in \mathcal{M}_d$, and at the corresponding vertex $(0, M) \in \widetilde{\mathcal{M}}_d$, are associated with the same vectors $e \in \mathbb{Z}^d \setminus \{0\}$.*

Proof. We first check that $(0, M) \in \widetilde{\mathcal{M}}_d$. Indeed, for any $e \in \mathbb{Z}^d \setminus \{0\}$, one has $\langle\!\langle l_e, (0, M)\rangle\!\rangle = \langle 0, e\rangle + \langle e, Me\rangle = \langle e, Me\rangle \geq 1$, since $M \in \mathcal{M}_d$.

We next prove that $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$, relying on the characterization of Remark 2.B.3. By assumption, since M is a vertex of $\mathcal{M}_d$, there exists $e_1, \cdots, e_I$ in $\mathbb{Z}^d \setminus \{0\}$ such that $\langle e_i, Me_i\rangle = 1$ for all $1 \leq i \leq I$, and $\mathrm{Span}\{e_i e_i^\top\}_{1\leq i\leq I} = S_d$. The latter property implies that $\{e_i\}_{i=1}^I$ spans $\mathbb{R}^d$, hence using (2.24) we obtain $\mathrm{Span}\{l_{e_i}\}_{1\leq |i|\leq I} = \mathbb{R}^d \times S^d$, with the usual convention $e_{-i} = -e_i$. Since $\langle\!\langle l_{\pm e_i}, (0, M)\rangle\!\rangle = \langle e_i, Me_i\rangle = 1$, we conclude that $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$. The additional point is straightforward, since the vectors $e \in \mathbb{Z}^d$ associated to active constraints at $(0, M) \in \widetilde{\mathcal{M}}_d$ are characterized by the identity $1 = \langle e, 0\rangle + \langle e, Me\rangle = \langle e, Me\rangle$. □

In the rest of this section, we compute the edges emanating from a vertex $(0, M) \in \widetilde{\mathcal{M}}_d$ in the form of Lemma 2.3.5. We apply the strategy of §2.B.3 to compute the outgoing direction of each edge, and eventually only encounter the two following cases:

(i) The computed edge direction has the form $\nu = (0, N)$ for some $N \in S_d$, hence the corresponding edge is internal to $\widetilde{\mathcal{M}}_d \cap (\{0\} \times S_d) = \{0\} \times \mathcal{M}_d$. Since the edges of $\mathcal{M}_d$ are known, see Corollary 2.2.11, this must be a bounded edge in the form of Theorem 2.3.1 (b).

(ii) The computed edge direction has the form $\nu = (v, vv^\top)$, where $v \in \mathbb{Z}^d$ (more precisely, v has the form indicated in Theorem 2.3.1 (c)). Thus for any e in $\mathbb{Z}^d \setminus \{0\}$,

$$\langle\!\langle l_e, \nu\rangle\!\rangle = \langle\!\langle (e, ee^\top), (v, vv^\top)\rangle\!\rangle = \langle e, v\rangle + \mathrm{Tr}(ee^\top vv^\top) = \langle e, v\rangle + \langle e, v\rangle^2.$$

Since e and v have integer coordinates, the scalar product $\langle e, v\rangle$ is an integer, and therefore $\langle\!\langle l_e, \nu\rangle\!\rangle \geq 0$ (with equality iff $\langle e, v\rangle \in \{0, -1\}$). Thus ν yields an unbounded edge in $\widetilde{\mathcal{M}}_d$ starting from $(0, M)$, in the form of Theorem 2.3.1 (c).

The graph defined by the edges and vertices of a regular polyhedron is connected, see Appendix 2.B. Once the above dichotomy is established, it follows that $\widetilde{\mathcal{M}}_d$ has no other vertices than those already found in Lemma 2.3.5, which concludes the proof of Theorem 2.3.1.

Notation (i-ii) and (A-D) in §2.3.1.1 and 2.3.1.2.

We establish the above dichotomy (i-ii) in §2.3.1.1 and 2.3.1.2, in dimension two and three respectively. For that purpose, we rely on the algorithm presented in §2.B.3 for enumerating the outgoing edges from a vertex in a polyhedron, and explicitly refer to its steps (A-D).

2.3.1.1 Edges of $\widetilde{\mathcal{M}}_2$

Let $b = (v_0, v_1, v_2)$ be a superbase of $\mathbb{Z}^2$, and let M_b be the corresponding vertex of $\mathcal{M}_2$, see (2.15). By Lemma 2.2.8, the active constraints at the vertex $M_b \in \mathcal{M}_2$ correspond to the set of vectors $E := \{e_{ij};\ i, j \in \{0, 1, 2\},\ i \neq j\}$ associated with the superbase b, see Definition 2.2.2. By Lemma 2.3.5, $(0, M_b)$ is a vertex of the polyhedron $\widetilde{\mathcal{M}}_2$, at which the constraints associated with the same vectors $e \in E$ are active. Since the number $\#(E) = 6$ of active constraints at $(0, M_b) \in \widetilde{\mathcal{M}}_2$ exceeds the dimension $\dim(\mathbb{R}^2 \times S_2) = 2 + 3 = 5$ of the embedding vector space, the vertex is degenerate. The edges containing $(0, M_b) \in \widetilde{\mathcal{M}}_2$ are obtained by selecting 4 out of the six active constraints, in other words by removing two elements from the set E. The following cases can be distinguished:

- Removing e_{12} and e_{21}. The corresponding direction is $\nu = (0, v_1 \otimes v_2)$, which lies within $\{0\} \times S_d$, and thus falls in case (i). Validation of the direction: one has $\langle\!\langle l_{e_{01}}, \nu \rangle\!\rangle = \langle e_{01}, v_1 \otimes v_2\, e_{01}\rangle = \langle e_{01}, v_1\rangle\langle e_{01}, v_2\rangle = 0$, since $\langle e_{01}, v_2\rangle = 0$. Likewise $\langle\!\langle l_e, \nu\rangle\!\rangle = 0$ for all $e \in \{\pm e_{01}, \pm e_{02}\}$, hence ν obeys the conditions of (B) of Algorithm §2.B.3.

- Removing e_{01} and e_{02}. The corresponding direction is $\nu = (v_0, v_0 \otimes v_0)$, which falls in the case (ii) of an unbounded edge. Validation of the direction: one has $\langle\!\langle l_{e_{12}}, \nu\rangle\!\rangle = \langle e_{12}, v_0\rangle^2 + \langle e_{12}, v_0\rangle = 0^2 + 0 = 0$, and $\langle\!\langle l_{e_{10}}, \nu\rangle\!\rangle = \langle e_{10}, v_0\rangle^2 + \langle e_{10}, v_0\rangle = (-1)^2 + (-1) = 0$. Likewise for $e \in \{e_{21}, e_{20}\}$.

- Removing e_{01} and e_{20}. The corresponding direction is $\nu = (v_0, v_1 \otimes v_1 - v_2 \otimes v_2)$, but it does not correspond to an edge, since it is eliminated in step (C) of Algorithm §2.B.3. Indeed, noting that $\langle l_e, \nu\rangle = \langle e, v_1\rangle^2 - \langle e, v_2\rangle^2 + \langle e, v_0\rangle$ we obtain

$$\begin{aligned} \langle l_{e_{10}}, \nu\rangle &= 1^2 - 0^2 - 1 = 0, \\ \langle l_{e_{02}}, \nu\rangle &= 0^2 - (-1)^2 + 1 = 0, \\ \langle l_{\pm e_{12}}, \nu\rangle &= (\pm 1)^2 - (\mp 1)^2 + 0, \end{aligned}$$

  showing that the direction ν is correct. However since

$$\langle l_{e_{01}}, \nu\rangle = (-1)^2 - 0^2 + 1 = 2, \qquad \langle l_{e_{20}}, \nu\rangle = 0^2 - 1^2 - 1 = -2,$$

  have opposite signs, the direction ν does not yield an edge of positive length.

There are 15 distinct two element subsets of $E := \{e_{ij};\ i, j \in \{0, 1, 2\},\ i \neq j\}$, and we have considered just three. However by permuting indices, the above considered cases respectively cover 3, 6, and again 6, distinct two element subsets E. Thus our enumeration is complete, and Theorem 2.3.1 is proved in dimension $d = 2$.

2.3.1.2 Edges of $\widetilde{\mathcal{M}}_3$

Let $b = (v_0, v_1, v_2, v_3)$ be a superbase of $\mathbb{Z}^3$, and let M_b be the corresponding vertex of $\mathcal{M}_3$, see (2.15). By Lemma 2.2.8, the active constraints at the vertex $M_b \in \mathcal{M}_3$ correspond to the set of vectors $E := \{e_{ij};\ i, j \in \{0, 1, 2, 3\},\ i \neq j\}$ associated with the superbase b, see Definition 2.2.2. By Lemma 2.3.5, $(0, M_b)$ is a vertex of the polyhedron $\widetilde{\mathcal{M}}_3$, at which the constraints associated with the same vectors $e \in E$ are active. Since the number $\#(E) = 12$ of active constraints at $(0, M_b) \in \widetilde{\mathcal{M}}_2$ exceeds the dimension $\dim(\mathbb{R}^3 \times S_3) = 3 + 6 = 9$ of the embedding vector space, the vertex is degenerate. The edges containing $(0, M_b) \in \widetilde{\mathcal{M}}_3$ are obtained by selecting 8 out of the twelve active constraints, in other words by removing four elements from the set E. The following cases can be distinguished:

- Removing $\pm e_{01}$ and two other unspecified elements of E. If the subset is not rejected in step (B), then the corresponding direction is $\nu = (0, v_0 \otimes v_1)$, which lies within $\{0\} \times S_d$ and thus falls into case (i). Validation of the direction: one has $\langle\!\langle l_{e_{ij}}, \nu \rangle\!\rangle = \langle e_{ij}, v_0\rangle\langle e_{ij}, v_1\rangle = 0$ as soon as $\{i,j\} \neq \{0,1\}$, hence ν obeys the conditions of (B).

- Removing $\alpha_{01}e_{01}$, $\alpha_{02}e_{02}$, $\alpha_{03}e_{03}$ and another unspecified element of E, where $\alpha_{01}, \alpha_{02}, \alpha_{03} \in \{-1, 1\}$. The corresponding direction is, up to a global sign change,

$$\nu = (-v_0,\ \alpha_{01}v_0 \otimes v_1 + \alpha_{02}v_0 \otimes v_2 + \alpha_{03}v_0 \otimes v_3).$$

  It is rejected in step (B) or (C), unless $\alpha_{01} = \alpha_{02} = \alpha_{03}$ in which case $\nu = (-\alpha_{01}v_0,\ v_0 \otimes v_0)$ (here with the correct sign) falls in case (ii) and defines an unbounded edge. (Note that $v_0 \otimes v_1 + v_0 \otimes v_2 + v_0 \otimes v_3 = -v_0 \otimes v_0$ since $v_1 + v_2 + v_3 = -v_0$.) Indeed, for any $i, j \in \{1, 2, 3\}$ such that $i \neq j$ one computes

$$\langle l_{\pm e_{0i}}, \nu\rangle = -\alpha_{0i} \mp 1, \qquad \langle l_{e_{ij}}, \nu\rangle = 0. \tag{2.26}$$

- Removing $\alpha_{01}e_{01}$, $-\alpha_{12}e_{12}$, $\alpha_{23}e_{23}$, $-\alpha_{30}e_{30}$, where $\alpha_{01}, \alpha_{12}, \alpha_{23}, \alpha_{30}$ belong to $\{-1, 1\}$. Then the corresponding direction is, up to a global sign change,

$$\nu = (v_1 + v_3,\ \alpha_{01}v_0 \otimes v_1 + \alpha_{12}v_1 \otimes v_2 + \alpha_{23}v_2 \otimes v_3 + \alpha_{30}v_3 \otimes v_0).$$

  It is rejected in step C, unless $\alpha_{01} = \alpha_{12} = \alpha_{23} = \alpha_{30}$, in which case $\nu = (v, v \otimes v)$ (here with the correct sign) with $v = -\alpha_{01}(v_1 + v_3) = \alpha_{01}(v_0 + v_2)$ falls in case (ii) and thus defines an unbounded edge. (Note that $-(v_1 + v_3) \otimes (v_1 + v_3) = (v_0 + v_2) \otimes (v_1 + v_3) = v_0 \otimes v_1 + v_1 \otimes v_2 + v_2 \otimes v_3 + v_3 \otimes v_0$ since $v_0 + v_2 = -(v_1 + v_3)$.) Indeed, we check that

$$\begin{aligned} \langle l_{e_{02}}, \nu\rangle &= 0 + 0 + 0 + 0 + 0, \quad \text{likewise for } e \in \{\pm e_{02}, \pm e_{13}\} \\ \langle l_{\pm e_{01}}, \nu\rangle &= -\alpha_{01} \pm 1, \quad \text{likewise for } e \in \{\pm e_{01}, \pm e_{12}, \pm e_{23}, \pm e_{30}\}. \end{aligned}$$

Finally, we need to show that all the possible 4 element subsets $S \subset \{e_{ij};\ i \neq j\}$ correspond to one of the considered cases, up to a permutation of the superbase. We refer to i and j as the indices of a vector $e_{ij} \in S$. If two elements of S share the same two indices, a.k.a. $e_{ij}, e_{ji} \in S$ for some $i \neq j$, then we fall in the first case. Otherwise, if (at least) three elements of S share one index, then we fall in the second case. Otherwise, each index $i \in \{0, \cdots, 3\}$ appears in at most two elements of S, thus exactly two since $\#(S) = 4 = \#\{0, \cdots, 3\}$. It follows that the indices of S define a cycle, and we fall in the last case.

2.3.2 A variant of Voronoi's first reduction

We introduce and study a variant of Voronoi's first reduction, applying to pairs (ω, D) of a vector $\omega \in \mathbb{R}^d$ and a positive definite symmetric tensor $D \in S_d^{++}$, instead of the matrix D alone in the original formulation (2.19). It is defined as follows:

$$\widetilde{\mathrm{Vor}}(\omega, D) := \inf\{\langle \omega, \eta\rangle + \mathrm{Tr}(DM);\ (\eta, M) \in \widetilde{\mathcal{M}}_d\}. \tag{2.27}$$

Somewhat surprisingly, this generalization of Voronoi's first reduction reduces to the original one, subject to a compatibility condition.

Theorem 2.3.6. *Let $d \leq 3$. For any $(D, \omega) \in S_d^{++} \times \mathbb{R}^d$ one has*

$$\widetilde{\mathrm{Vor}}(D, \omega) = \begin{cases} -\infty & \text{if } \exists v \in \mathbb{Z}^d \setminus \{0\},\ \langle v, Dv\rangle + \langle \omega, v\rangle < 0, \\ \mathrm{Vor}(D) & \text{otherwise.} \end{cases}$$

Proof. The result follows from the description of the vertices and unbounded edges of the polytope $\widetilde{\mathcal{M}_d}$ in Theorem 2.3.1, and from the general expression (2.41) of the value of a linear program. Note also that any $v_1 \in \mathbb{Z}^d \setminus \{0\}$ with co-prime coordinates can be completed into a basis $v_1, \cdots, v_d$ of $\mathbb{Z}$, hence also into a superbase with $v_0 := -(v_1 + \cdots + v_d)$. Hence the set of directions of all unbounded edges of $\widetilde{\mathcal{M}_d}$, see Theorem 2.3.1 (c), is $\mathbb{Z}^d \setminus \{0\}$. □

Proposition 2.3.7. *Let $d \leq 3$, and let $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$. The following are equivalent:*

(i) The pair (ω, D) is absolutely feasible.

(ii) The linear program $\widetilde{\mathrm{Vor}}(\omega, D)$ is bounded.

In case (ii), any set of KKT relations for $\widetilde{\mathrm{Vor}}(\omega, D)$ yields a simultaneous decomposition of (ω, D), showing (i) explicitly.

Proof. Proof that (i) ⇒ (ii). Assume that (ω, D) is absolutely feasible, and denote by $\rho_i \geq 0$ the weights, and $e_i \in \mathbb{Z}^d$ the offsets of the corresponding decomposition, so that

$$\omega = \sum_{1 \leq i \leq I} \rho_i e_i, \qquad\qquad D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^T. \tag{2.28}$$

Then for any $(M, \eta) \in \widetilde{\mathcal{M}_d}$, one obtains using the identity $\langle e, Me \rangle = \mathrm{Tr}(Mee^\top)$

$$\langle \omega, \eta \rangle + \mathrm{Tr}(DM) = \sum_{1 \leq i \leq I} \rho_i \left(\langle \omega, e_i \rangle + \langle e_i, Me_i \rangle \right) \geq \sum_{1 \leq i \leq I} \rho_i \geq 0.$$

Therefore $\widetilde{\mathrm{Vor}}(D, \omega) \geq 0 > -\infty$ is bounded.

Proof that (ii) ⇒ (i). By Proposition 2.2.12 there exists a vertex M_b of $\mathcal{M}_d$, where b is a superbase of $\mathbb{Z}^d$, such that $\mathrm{Vor}(D) = \mathrm{Tr}(DM_b)$. By Lemmas 2.2.8 and 2.3.5, $(0, M_b)$ is a vertex of $\widetilde{\mathcal{M}_d}$ at which finitely many constraints $(e_i)_{i=1}^I$ are active. By Theorem 2.3.6, $\widetilde{\mathrm{Vor}}(D) = \mathrm{Vor}(D) = \mathrm{Tr}(DM_b) = \langle \omega, 0 \rangle + \mathrm{Tr}(DM_b)$ and this minimum is attained at the vertex $(0, M_b)$. The KKT relations express that there exists non-negative weights $(\rho_i)_{i=1}^I$ (possibly non-unique) such that the objective function and the weighted sum of the constraints are equal: one has $\langle \omega, \eta \rangle + \mathrm{Tr}(DM) = \sum_{1 \leq i \leq I} \rho_i (\langle e_i, \eta \rangle + \langle e_i, Me_i \rangle)$ for all $(\eta, M) \in \mathbb{R}^d \times S_d$. From this point, the simultaneous decomposition (2.28) follows by identification, as announced. □

Remark 2.3.8 (Degeneracy of the vertices of $\widetilde{\mathcal{M}_d}$). The vertices of $\widetilde{\mathcal{M}_d}$ are degenerate, in dimension $d \in \{2, 3\}$, in the sense that exactly $d(d+1)$ constraints are active, which is strictly greater than $\dim(\mathbb{R}^d \times S_d) = d(d+1)/2 + d$. As a result, the KKT relations for the linear program $\mathrm{Vor}(\omega, D)$ in general do not *uniquely* determine the decomposition (2.28) of the pair (ω, D). This is in contrast with Voronoi's first reduction in dimension $d \leq 3$, see Remark 2.2.13.

2.3.3 Local study of feasibility

In this section, we compare the conditions of canonical and absolute feasibility of a pair (ω, D), in dimension $d \leq 3$, concluding the proof of Theorem 2.1.6. For that purpose, we fix a symmetric positive definite matrix $D \in S_d^{++}$, denote by $b = (v_0, \cdots, v_d)$ a D-obtuse superbase, and recall Selling's decomposition (2.13)

$$D = \sum_{0 \leq i < j \leq d} \sigma_{ij} e_{ij} e_{ij}^\top, \tag{2.29}$$

where $\sigma_{ij} := -\langle v_i, Dv_j\rangle \geq 0$ and where $e_{ij} \in \mathbb{Z}^d \setminus \{0\}$ for all $0 \leq i < j \leq d$ is introduced in Definition 2.2.2. In this subsection, for notational convenience, the indices i and j, are always implicitly constrained to lie in the set $\{0, \cdots, d\}$.

We characterize, in the next proposition, the canonical and absolute feasibility of a pair (ω, D) in terms of Selling's decomposition of D. The argument, in the case of absolute feasibility, heavily relies on the results established in §2.3.2. An interesting byproduct is that, if a pair (ω, D) is absolutely feasible, then it admits a decomposition (2.3) whose offsets are those of Selling's formula (2.29) for D (and the opposite offsets).

Proposition 2.3.9. *Assume $d \leq 3$. Let $\omega \in \mathbb{R}^d$ and $D \in S_d^{++}$. We use the notations b and $(\sigma_{ij}, e_{ij})_{i<j}$ of Selling's decomposition* (2.29)*. Then*

- *(ω, D) is absolutely feasible iff there exists $\mu_{ij} \in [-1, 1]$, for all $0 \leq i < j \leq d$, such that $\omega = \sum_{i<j} \mu_{ij}\sigma_{ij}e_{ij}$.*
- *(ω, D) is canonically feasible iff $|\langle e_{ij}, D^{-1}\omega\rangle| \leq 1$ for all $0 \leq i < j \leq d$ such that $\sigma_{ij} > 0$.*

Proof. First equivalence. If the pair (ω, D) is absolutely feasible, then by Proposition 2.3.7 the linear program $\widetilde{\mathrm{Vor}}(\omega, D)$ is bounded, and attains its minimum at the vertex $(0, M_b)$, at which the active constraints are associated with the vectors e_{ij}, $i \neq j$. By the KKT relations, there exists non-negative weights ρ_{ij}, $i \neq j$, such that

$$\omega = \sum_{i\neq j} \rho_{ij} e_{ij} \qquad\qquad D = \sum_{i\neq j} \rho_{ij} e_{ij} e_{ij}^T.$$

Recalling that $e_{ji} = -e_{ij}$ for all $i \neq j$, we obtain

$$\omega = \sum_{i<j} (\rho_{ij} - \rho_{ji}) e_{ij} \qquad\qquad D = \sum_{i<j} (\rho_{ij} + \rho_{ji}) e_{ij} e_{ij}^T.$$

By uniqueness of Selling's decomposition, see Remark 2.2.13, one has $\sigma_{ij} = \rho_{ij} + \rho_{ji}$ for all $i < j$. Denoting $\mu_{ij} := (\rho_{ij} - \rho_{ji})/\sigma_{ij} \in [-1, 1]$ when $\sigma_{ij} > 0$ (and e.g. $\mu_{ij} = 0$ if $\sigma_{ij} = 0$), we obtain $\omega = \sum_{i<j} \mu_{ij}\sigma_{ij}e_{ij}$ as announced. The reverse implication is trivial, by defining $\rho_{ij} = \sigma_{ij}(1 + \mu_{ij})/2$ and $\rho_{ji} = \sigma_{ij}(1 - \mu_{ij})/2$, for all $i < j$.

Second equivalence. By construction, see Definition 2.1.5, the pair (ω, D) obeys is canonically feasible iff $\sigma_{ij}(1 + \varepsilon\langle e_{ij}, \eta\rangle) \geq 0$ for all $i < j$ and all $\varepsilon \in \{-1, 1\}$, where $\eta := D^{-1}\omega$. This is indeed equivalent to $|\langle e_{ij}, D^{-1}\omega\rangle| \leq 1$, for all $i < j$ such that $\sigma_{ij} > 0$, as announced. □

We next state two technical lemmas which, combined with Proposition 2.3.9 above, let us conclude the proof of Theorem 2.1.6. The proof of Lemma 2.3.10 is postponed to §2.3.3.1.

Lemma 2.3.10. *Let $D \in S_d^{++}$, $d \leq 3$. We use the notations b, $(\sigma_{ij}, e_{ij})_{i<j}$ of Selling's decomposition* (2.29)*. Then $|\langle e_{ij}, Me_{kl}\rangle| \leq \|e_{ij}\|_M^2$ for all $i < j$ and all $k < l$, where $M := D^{-1}$.*

Lemma 2.3.11. *Let $D = \sum_{r=1}^R \sigma_r e_r e_r^T$, where $\sigma_r \geq 0$, $e_r \in \mathbb{R}^d$ for all $1 \leq r \leq R$, and R is a positive integer. If D is positive definite, then $\sigma_r\langle e_r, D^{-1}e_r\rangle \leq 1$ for all $1 \leq r \leq R$.*

Proof. For any $1 \leq r \leq R$, one has $D \succeq \sigma_r e_r e_r^\top$, in the sense of symmetric matrices. Therefore, letting $v_r := D^{-1}e_r$, we obtain $\langle e_r, D^{-1}e_r\rangle = \langle v_r, Dv_r\rangle \geq \sigma_r\langle v_r, e_r\rangle^2 = \sigma_r\langle e_r, D^{-1}e_r\rangle^2$. This implies $1 \geq \sigma_r\langle e_r, D^{-1}e_r\rangle$, as announced. □

Proof of Theorem 2.1.6. Assume that (D, ω) is absolutely feasible. Then for any $i < j$, using the notations of Proposition 2.3.9, we obtain

$$\begin{aligned} |\langle e_{ij}, D^{-1}\omega\rangle| &\leq \sum_{k<l} \sigma_{kl}|\mu_{kl}|\,|\langle e_{ij}, D^{-1}e_{kl}\rangle| \\ &\leq \sum_{k<l} \sigma_{kl}\langle e_{kl}, D^{-1}e_{kl}\rangle \\ &\leq \sum_{k<l} 1 = d(d+1)/2. \end{aligned}$$

The three inequalities follows, successively, from Proposition 2.3.9 (first point), Lemma 2.3.10, and Lemma 2.3.11. If follows from Proposition 2.3.9 (second point) that $(D, \omega/C)$ is canonically feasible, with $C := d(d+1)/2$, as announced. □

2.3.3.1 Proof of Lemma 2.3.10

Throughout this subsection, we use for convenience the notation $\langle v, w\rangle_M := \langle v, Mw\rangle$, for any $v, w \in \mathbb{R}^d$, $M \in S_d^{++}$. We use the notations of Lemma 2.3.10. In particular $D \in S_d^{++}$, $M := D^{-1}$, $b = (v_0, \cdots, v_d)$ is a D-obtuse superbase, and $(\sigma_{ij}, e_{ij})_{i<j}$ are the coefficients and offsets of Selling's decomposition (2.29) of D. As before, the indices i, j implicitly lie in $\{0, \cdots, d\}$.

Proof in dimension $d = 2$. Assume that the superbase $b = (v_0, v_1, v_2)$ satisfies

$$\det(v_1, v_2) = 1,$$

without loss of generality and up to exchanging v_1 and v_2. Then $(e_{12}, e_{20}, e_{01}) = (v_0^\perp, v_1^\perp, v_2^\perp)$ by (2.11), and this triplet is an M-obtuse superbase by Lemma 2.2.17. Denoting $(w_0, w_1, w_2) := (e_{12}, e_{20}, e_{01})$ one obtains

$$-\langle w_0, w_1\rangle_M - \langle w_0, w_2\rangle_M = \langle w_0, -w_1 - w_2\rangle_M = \|w_0\|_M^2,$$

and therefore $0 \leq -\langle w_0, w_1\rangle_M \leq \|w_0\|_M^2$. Likewise $0 \leq -\langle w_i, w_j\rangle_M \leq \|w_i\|_M^2$ for all $i \neq j$, which is the announced result. □

Proof in dimension $d = 3$. Denote $w_{ij} := v_i \times v_j$ for all $i \neq j$. In the following, $\{i, j, k, l\}$ denotes an arbitrary permutation of $\{0, 1, 2, 3\}$, thus for instance $w_{ij} = \pm e_{kl}$ by (2.11). Note also that

$$w_{ij} = -w_{ji}, \qquad \text{and} \qquad w_{ij} + w_{ik} + w_{il} = v_i \times (v_j + v_k + v_l) = -v_i \times v_i = 0.$$

The scalar products defined by $D \in S_3^{++}$ and its inverse $M := D^{-1}$ are related by the following identity, where $u, v, w \in \mathbb{R}^3$

$$\det(D)\langle u \times v,\, u \times w\rangle_M = \|u\|_D^2\langle v, w\rangle_D - \langle u, v\rangle_D\langle u, w\rangle_D.$$

(In the case $D = \mathrm{Id}$ this is known as the Binet-Cauchy identity. In the general case where $D = A^T A$ for some $A \in \mathrm{GL}_3(\mathbb{R})$ it follows from a linear change of variables and the relation $(Au) \times (Av) = \mathrm{cof}(A)(u \times v)$ where $\mathrm{cof}(A)$ denotes the cofactor matrix.)

Choosing $u = v_i$, $v = v_j$ and $w = v_k$, we obtain that $\langle w_{ij}, w_{ik}\rangle_M \leq 0$. On the other hand

$$\begin{aligned} -\langle w_{ij}, w_{ik}\rangle_M - \langle w_{ij}, w_{il}\rangle_M &= \langle w_{ij}, v_i \times (-v_k - v_l)\rangle_M \\ &= \langle w_{ij}, v_i \times (v_i + v_j)\rangle_M \end{aligned}$$

$$= \|w_{ij}\|_M^2,$$

thus $0 \leq -\langle w_{ij}, w_{ik}\rangle_M \leq \|w_{ij}\|_M^2$. Finally, since $-w_{kl} = w_{ki} + w_{kj}$, we obtain that

$$-\langle w_{ij}, w_{kl}\rangle_M = \langle w_{ij}, w_{ki} + w_{kj}\rangle_M = -\langle w_{ij}, w_{ik}\rangle_M + \langle w_{ji}, w_{jk}\rangle_M,$$

and therefore, by the previous estimate, $-\|w_{ij}\|_M^2 \leq \langle w_{ij}, w_{kl}\rangle_M \leq \|w_{ij}\|_M^2$. This concludes the proof of Lemma 2.3.10. □

2.4 Numerical experiments

We illustrate the PDE discretization introduced in this paper with synthetic numerical experiments, in dimension $d \in \{2, 3\}$, involving linear and semi-linear operators, and using Dirichlet boundary conditions on a non-square and non-smooth domain. Let us mention that a close variant of the proposed scheme, involving the divergence form operator $\mathrm{div}(D(x)(\nabla u(x) - \omega(x)u(x)))$ featuring both a first and second order term, is used in [Par+19] for image inpainting purposes in dimension $d = 2$, in collaboration with Jean-Marie Mirebeau. See also [FM14] for applications to image denoising in dimension $d \in \{2, 3\}$, with an operator lacking the first order term, however. Additional concrete applications of the proposed scheme will be the object of future work.

The PDEs addressed numerically in this section take the form

$$\mathcal{L}u(x) = f(x),\ \forall x \in \Omega, \qquad\qquad u(x) = g(x),\ \forall x \in \partial\Omega, \tag{2.30}$$

where $\Omega := \{x \in \mathbb{R}^d;\ \|x\| < 1\} \cup]0,1[^d$ is the union of the d-dimensional unit ball and of the d-dimensional unit cube. The PDE operator $-\mathcal{L}u(x)$ is chosen as the following linear (resp. semi-linear) expression

$$\langle \omega(x), \nabla u(x)\rangle + \tfrac{1}{2}\,\mathrm{Tr}(D(x)\nabla^2 u(x)) \quad \Big(\text{resp. } \tfrac{1}{2}\langle \omega(x), \nabla u(x)\rangle^2 + \tfrac{1}{2}\,\mathrm{Tr}(D(x)\nabla^2 u(x))\Big) \tag{2.31}$$

whose coefficients $\omega\colon \overline{\Omega} \to \mathbb{R}^d$ and $D\colon \overline{\Omega} \to S_d^{++}$ are defined for any $x = (x_1, \cdots, x_d)$ in $\mathbb{R}^d$ by

$$\begin{aligned} \omega(x) &:= \frac{2 - \cos(\pi x_1)}{3}\omega_0(x), \\ D(x) &:= \mu\frac{2 + \cos(\pi x_1)}{3}\big(\nu I_d + (1-\nu)\omega_0(x/2)\omega_0(x/2)^T\big), \end{aligned} \tag{2.32}$$

where the parameters $\mu, \nu > 0$ are specified in Figures 2.3 to 2.6, and where ω_0 is the field of unit vectors defined by

$$\omega_0(x) := \begin{cases} (\cos(\pi x_2), \sin(\pi x_2)) & \text{if } d = 2, \\ (\cos(\pi x_2), \sin(\pi x_2)\cos(\pi x_3), \sin(\pi x_2)\sin(\pi x_3)) & \text{if } d = 3. \end{cases}$$

This particular choice of operator and coefficients is only meant to be reasonably simple and explicit, and to feature substantial anisotropy for the second order term. It also allows for a direct analytic verification of the assumptions of Theorem 2.1.7 ensuring the DDE property, see the last paragraph of this section.

By construction, the condition number $\mathrm{Cond}(D) := \sqrt{\|D\|\|D^{-1}\|}$ of the symmetric matrices involved in the PDE operators (2.31) is $\mathrm{Cond}(D(x)) = \nu^{-\frac{1}{2}}$. In the experiments we use $\nu = 1/10$, so that $\mathrm{Cond}(D(x)) = \sqrt{10} \approx 3.1$, and the radius of the numerical scheme stencil never exceeds $\sqrt{5}$ (times the discretization grid scale h) both in dimension two and three. Stronger

Stencil radius	$d=2$	$d=3$	$d=3$, $\lambda_1=\lambda_2\leq\lambda_3$
$\mathrm{Cond}(D)\leq\sqrt{10}$	$\sqrt{5}\approx 2.2$	$\sqrt{11}\approx 3.3$	$\sqrt{5}\approx 2.2$
$\mathrm{Cond}(D)\leq 10$	$\sqrt{26}\approx 5.1$	$\sqrt{69}\approx 8.3$	$\sqrt{30}\approx 5.5$

Table 2.1: Maximum stencil radius of the proposed discretization, for a unit grid scale, depending on the condition number of D defining the second order term. In other words $\max_{1\leq i\leq I}\|e_i\|$ where $D=\sum_{1\leq i\leq I}\sigma_i e_i e_i^T$ is Selling's decomposition, see Lemma 2.2.3. First column : $D\in S_2^{++}$, second column : $D\in S_3^{++}$, third column : $D\in S_3^{++}$ with the two smallest eigenvalues equal, as in (2.32). Values obtained experimentally using a fine sampling of the corresponding sets of matrices, see Theorem 2.1.9 for a (non-sharp) proved upper bound of the form $C_d\,\mathrm{Cond}(D)$.

anisotropies, up to $\mathrm{Cond}(D(x))=10$, are routinely used in the applications of numerical schemes based on Selling's algorithm ([BCM16; BOZ04; FM14] and Chapter 5) and in particular the approximation of sub-Riemannian and non-holonomic eikonal equations [Mir18; Mir19]. However, this approach looses relevance for even stronger anisotropies $\mathrm{Cond}(D(x))\gg 10$, because the numerical scheme stencils become excessively wide and accuracy therefore degrades. In this case one may limit the size of the stencils at the price of a consistency error [BOZ04, section 6], or alternatively switch to completely different techniques such as asymptotic preserving schemes [Deg+12].

For any discretization step $h>0$, we let $\Omega_h:=\Omega\cap h\mathbb{Z}^d$ and consider the finite differences scheme

$$L_h u(x)=f(x)\quad\text{in }\Omega_h, \tag{2.33}$$

where one has, denoting $g(x,p):=\langle\omega(x),p\rangle$ in the linear case (resp. $g(x,p):=\frac{1}{2}\langle\omega(x),p\rangle^2$ in the semi-linear case)

$$-L_h u(x):=g(x,D(x)^{-1}\nabla_h^{D(x)}u(x))+\tfrac{1}{2}\Delta_h^{D(x)}u(x).$$

The Dirichlet boundary condition from (2.30) does not appear in (2.33) because it is implicitly implemented via the finite differences operators, defined as (2.37) and (2.38) when the point x is near $\partial\Omega$. See Appendix 2.A for more discussion on the extension of the scheme of Definition 2.1.5 to non-constant coefficients, Dirichlet boundary conditions, and non-linear operators.

As announced, we present synthetic tests of our numerical scheme. For that purpose, a function $u\colon\overline{\Omega}\to\mathbb{R}$ is chosen with a closed form expression, and the right-hand side $f\colon\Omega\to\mathbb{R}$ is generated by symbolic differentiation and evaluation of $\mathcal{L}u$, so that u obeys (2.30) with boundary condition $g:=u_{|\partial\Omega}$. The discretized PDE (2.33) is then solved for a range of grid scales $h>0$, and the resulting $l^1(\Omega_h)$ and $l^\infty(\Omega_h)$ reconstruction errors are reported in Figures 2.3 to 2.6.

The chosen exact solutions are a smooth function $\mathbf{u}_1$, a $C^{2,0.5}$ function $\mathbf{u}_2$, and a singular function $\mathbf{u}_3$, inspired by [FJ17] for $\mathbf{u}_1$ and by [FO13] for $\mathbf{u}_2$ and $\mathbf{u}_3$, and defined in $\overline{\Omega}$ by

$$\mathbf{u}_1(x):=\frac{1}{4}\|x\|^4,\qquad \mathbf{u}_2(x):=\max(0,\|x\|-0.4)^{2.5},\qquad \mathbf{u}_3(x):=\sqrt{d-\|x\|^2}. \tag{2.34}$$

The multiplicative coefficient $1/4$ in the definition of $\mathbf{u}_1$ is chosen so that the range of values taken by $\|\nabla\mathbf{u}_1\|$ in Ω remains close to the one of values taken by $\|\nabla\mathbf{u}_2\|$, since the gradient magnitude influences the DDE property of the scheme (2.33) in the semi-linear case, see section 2.4.1. In numerical experiments, we also adjust the parameter μ in the definition of the tensor field $D\colon\overline{\Omega}\to S_d^{++}$ so that DDE holds at reasonable grid scales.

Empirically we observe second order convergence $\|\mathbf{u}-u_h\|_1=\mathcal{O}(h^2)$ and $\|\mathbf{u}-u_h\|_\infty=\mathcal{O}(h^2)$, where $\mathbf{u}$ is among the two test functions $\mathbf{u}_1$ and $\mathbf{u}_2$ defined in (2.34) and u_h is the numerical

solution of (2.33) with the corresponding r.h.s. for both the linear and semi-linear operators (2.31), in both dimension two and three, see Figures 2.3 to 2.6. For the test function $\mathbf{u}_3$, first order convergence $\|\mathbf{u}_3 - u_h\|_1 = \mathcal{O}(h)$ and $\|\mathbf{u}_3 - u_h\|_\infty = \mathcal{O}(h)$ is observed instead. From a theoretical standpoint, convergence was not expected for $\mathbf{u}_3$ and the semi-linear scheme, since the DDE property is not guaranteed in this case, even for small h.

For the semi-linear equations, a Newton method is used, converging in at most 12 iterations in our experiments with tolerance 10^{-8} on the max-norm of residual of the discretized PDE.

Remark 2.4.1 (Dominant source of numerical error). The curves of convergence associated to the *linear* and *semi-linear* equations are conspicuously similar in several cases: for the function $\mathbf{u}_2$ in dimension two, see Figures 2.3 and 2.4, and for $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ in dimension three, see Figures 2.5 and 2.6. This suggests that the discretization of the first order term in (2.31) is not the dominant source of error in these cases.

For $\mathbf{u}_1$ and $\mathbf{u}_3$ in dimension three, we obtained a different convergence curve when changing the tensor field $D : \overline{\Omega} \to S_d^{++}$, suggesting that the discretization of $\mathrm{Tr}(D\nabla^2 u)$ is the dominant source of numerical error. For the $C^{2,0.5}$ function $\mathbf{u}_2$, we did *not* observe a significant difference in the curves of convergence when changing the tensor field D, but we did observe one when replacing the radius $r = 0.4$ with 0.5 in its definition, suggesting that the dominant source of error is related to the configuration of the grid points Ω_h in the vicinity of the sphere of radius r across which $\mathbf{u}_2$ is non-smooth.

2.4.1 Theoretical guarantees of Discrete Degenerate Ellipticity

An a priori analysis allows to guarantee the DDE property of the numerical schemes used in our numerical experiments (except in one case where it fails), thanks to the explicit and reasonably simple expression of the PDE coefficients (2.31) (and, in the semi-linear case, of the PDE solution (2.34)). In practical applications, such an analysis may not be possible, but alternatively the DDE property can be checked numerically by looking at the sign of the coefficients of the Jacobian matrix of the discretized operator L_h.

Letting $M(x) := D(x)^{-1}$, one easily obtains

$$\|M(x)\| = \mu^{-1}(3/(2 + \cos(\pi x_1)))\nu^{-1} \leq 3\mu^{-1}\nu^{-1},$$

and therefore

$$\|M(x)\|^{1/2}\|\omega(x)\|_{M(x)} \leq \|M(x)\|\|\omega(x)\| \leq 3\mu^{-1}\nu^{-1}.$$

It follows that the pair $(h\omega(x), D(x))$ is canonically feasible as soon as $h \leq c_d\mu\nu/3$, where the absolute constant c_d is specified in Theorem 2.1.7. The discretization of the linear operator (2.31, left) is thus DDE under these conditions.

We now check whether the discretization of the *semi*-linear operator (2.31, right) is DDE in a neighborhood of the solutions (2.34), by linearizing the operator. For any $x \in \overline{\Omega}$ and $p \in \mathbb{R}^d$ one has $\|\nabla_p g(x, \nabla u(x))\| = \|\langle\omega(x), \nabla u(x)\rangle\omega(x)\| \leq \|\omega(x)\|^2\|\nabla u(x)\| \leq \|\nabla u(x)\|$. By the same reasoning as above, if $\mathbf{u}$ denotes either one of the functions $\mathbf{u}_1$ and $\mathbf{u}_2$ in (2.34), then the pair $(h\nabla_p g(x, \nabla\mathbf{u}(x)), D(x))$ is canonically feasible for all $x \in \Omega$, and thus the scheme (2.33) is DDE in the neighborhood of $\mathbf{u}$, as soon as

$$h\|\nabla\mathbf{u}\|_{L^\infty(\Omega)} < c_d\mu\nu/3,$$

where we used that $\nabla\mathbf{u}_1$ and $\nabla\mathbf{u}_2$ are bounded on Ω. In contrast $\|\nabla\mathbf{u}_3(x)\|$ is unbounded when $x \to (1, \cdots, 1) \in \partial\Omega$. Thus DDE fails in the neighborhood of $\mathbf{u}_3$, but as noted above we do still observe convergence empirically in this particular case.

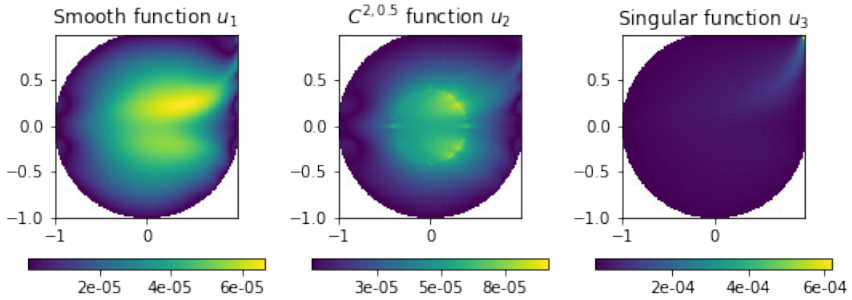


Figure 2.2: Errors in numerical solutions to the linear equation in dimension $d = 2$, with parameters $\mu = 2$, $\nu = 1/10$, $h = 1/100$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$.

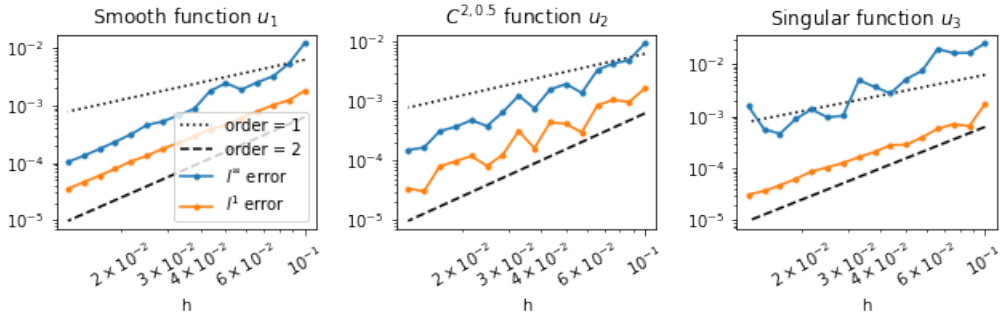


Figure 2.3: Convergence of the numerical scheme for the linear equation in dimension $d = 2$, with parameters $\mu = 2$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. Degenerate ellipticity is guaranteed by section 2.4.1 for $h \leq 1/30 \approx 0.0333$ and empirically observed up to $h \approx 0.0660$.

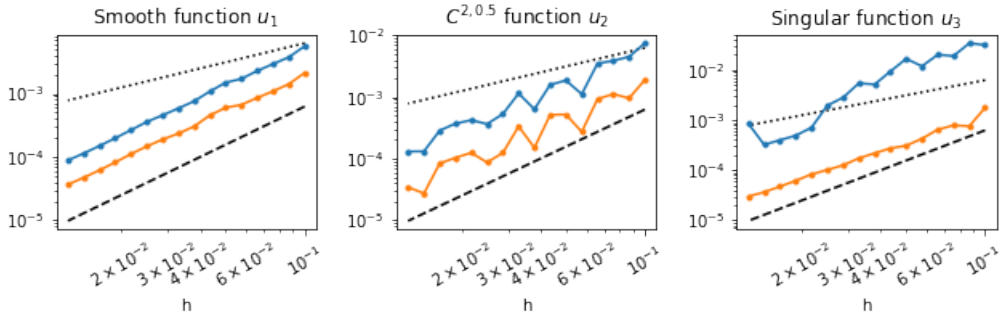


Figure 2.4: Convergence of the numerical scheme for the semi-linear equation in dimension $d = 2$, with parameters $\mu = 4$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 2.3. In the neighborhood of functions $\mathbf{u}_1$ and $\mathbf{u}_2$, degenerate ellipticity is guaranteed by section 2.4.1 respectively for $h < 1/(30\sqrt{2}) \approx 0.0236$ and for $h < 1/(75(\sqrt{2} - 0.4)^{1.5}) \approx 0.0131$. It is observed empirically in the last iteration of the Newton method respectively up to $h \approx 0.0379$ and up to $h \approx 0.0435$. In the case of the singular function $\mathbf{u}_3$, degenerate ellipticity is not theoretically guaranteed, but it is nevertheless observed empirically in the last iteration of the Newton method up to $h \approx 0.0574$.

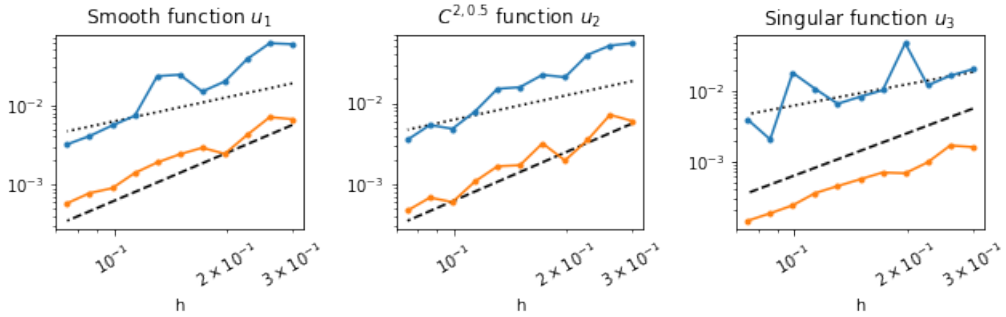


Figure 2.5: Convergence of the numerical scheme for the linear equation in dimension $d = 3$, with parameters $\mu = 8$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 2.3. Degenerate ellipticity is guaranteed by section 2.4.1 for $h \leq 1/(5\sqrt{3}) \approx 0.115$ and empirically observed up to $h \approx 0.198$.

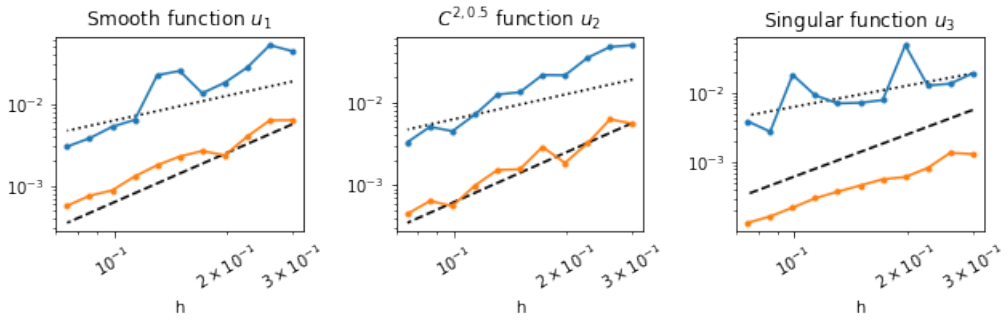


Figure 2.6: Convergence of the numerical scheme for the semi-linear equation in dimension $d = 3$, with parameters $\mu = 16$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 2.3. In the neighborhood of functions $\mathbf{u}_1$ and $\mathbf{u}_2$, degenerate ellipticity is guaranteed by section 2.4.1 respectively for $h < 2/135 \approx 0.0148$ and for $h < 1/(75\sqrt{3}(\sqrt{3}-0.4)^{1.5}) \approx 0.00501$. It is observed empirically in the last iteration of the Newton method respectively up to $h \approx 0.131$ and up to $h \approx 0.261$. In the case of the singular function $\mathbf{u}_3$, degenerate ellipticity is not theoretically guaranteed, but it is nevertheless observed empirically in the last iteration of the Newton method up to $h = 0.3$, that is, for all values of h we tested in the graphs above.

2.5 Conclusion and perspectives

In this paper, we answer whether one can discretize linear PDE operator, of order at most two and in dimension $d \leq 3$, using a second order consistent finite difference scheme obeying the discrete degenerate ellipticity property. The question is basic and of broad interest, and in dimension $d = 1$ the answer is indeed simple, well known, and taught at a basic level. In dimension $d \in \{2, 3\}$ however the anisotropy of the second order part of the operator comes into play, and a subtler analysis is required. Leveraging tools from the field of Euclidean lattice geometry, we could characterize whether a discretization exists, and provide an explicit (quasi-)optimal construction. Numerical experiments illustrate the efficiency of the method in dimension $d \in \{2, 3\}$, on linear and semi-linear problems.

Several research directions are open, both practical and theoretical, including (i) applications to PDEs arising from concrete problems, especially those whose first order term is large, e.g. depending on a relaxation parameter, (ii) extensions to fully non-linear HJB PDEs, and (iii) a theoretical analysis of the convergence rates. Another interesting open problem is the extension of our results for a dimension $d > 3$, which is not obvious, since a key ingredient of our analysis known as D-obtuse superbases does not necessarily exist in that case, see Definition 2.2.1 and the discussion below.

2.A Adaptation to semi-linear and fully non-linear PDEs

The numerical scheme presented in the introduction of this paper applies to *linear schemes* with *constant coefficients*, defined over the *full space* $\mathbb{R}^d$, $d \in \{2, 3\}$. We illustrate in this appendix how the three restrictions in emphasis can be relaxed. For that purpose let us recall the definition of a discrete degenerate elliptic scheme, in a general setting.

Definition 2.A.1. Let X be a discrete set, and for each $x \in X$ let $V(x) \subset X \setminus \{x\}$ be a finite set (the neighbors, or stencil of x). Let also $\mathbb{U} := \mathbb{R}^X$. A numerical scheme on X, with stencil V,

is a mapping $F : \mathbb{U} \to \mathbb{U}$ of the form

$$Fu(x) := \mathcal{F}(x, u(x), [u(x) - u(y)]_{y \in V(x)}).$$

It is said discrete degenerate elliptic (DDE) if $\mathcal{F}$ is non-decreasing w.r.t. the second and third arguments (coordinate wise).

Definition 2.1.1, from the introduction, is a special case of Definition 2.A.1, adapted to linear schemes with constant coefficients, and choosing $X = h\mathbb{Z}^d$ and $V(x) := \{x + he_i;\ 1 \leq |i| \leq I\}$. In the rest of this appendix, we show how various natural extensions of our numerical scheme fit into the general framework of Definition 2.A.1.

Non-constant coefficients

Discrete Degenerate Ellipticity is a local property, which only needs to be verified pointwise, independently at each point $x \in X$ of the discretization domain, see Definition 2.A.1. As a result, the numerical scheme presented in this paper trivially extends to non-constant coefficients. More precisely, let ω and D be a field of vectors and of symmetric positive definite matrices, and let $h > 0$ be a grid scale. Then we can define the counterparts with variable coefficients of the linear PDE operator (2.1) and of its canonical discretization (2.7)

$$-\mathcal{L}u(x) := \langle \omega(x), \nabla u(x) \rangle + \tfrac{1}{2} \operatorname{Tr}(D(x)\nabla^2 u(x)), \tag{2.35}$$

$$-L_h u(x) := \langle D(x)^{-1}\omega(x), \nabla_h^{D(x)} u(x) \rangle + \tfrac{1}{2}\Delta_h^{D(x)} u(x). \tag{2.36}$$

The scheme L_h is DDE under the same conditions, pointwise, as in the constant coefficient case. It is not hard to show that the coefficients $x \mapsto \rho_i(x) \geq 0$ of L_h expressed as in (2.1) are Lipschitz, provided ω and D are Lipschitz. Interestingly, convergence rates have been established in a similar setting [Kry05] but under the slightly stronger assumption that $x \mapsto \sqrt{\rho_i(x)}$ is Lipschitz.

Dirichlet boundary conditions

Consider a bounded open domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, equipped with Dirichlet boundary conditions $f : \partial\Omega \to \mathbb{R}$, and let $\Omega_h := \Omega \cap h\mathbb{Z}^d$, where the grid scale $h > 0$ is fixed in the following. For all $x \in \Omega_h$, $e \in \mathbb{Z}^d \setminus \{0\}$, define $h_x^e := \min\{k > 0;\ x + ke \in \Omega_h \cup \partial\Omega\}$, and note that $0 < h_x^e \leq h$. Introduce the first and second finite difference operators, where for convenience we denote $h^\pm := h_x^{\pm e}$, and where $u : \Omega_h \to \mathbb{R}$ is extended to $\partial\Omega$ using the provided Dirichlet boundary condition

$$\delta_h^e u(x) := \frac{1}{2}\Big(\frac{u(x + h^+ e) - u(x)}{h^+} - \frac{u(x - h^- e) - u(x)}{h^-}\Big), \tag{2.37}$$

$$\Delta_h^e u(x) := \frac{2}{h^+ + h^-}\Big(\frac{u(x + h^+ e) - u(x)}{h^+} + \frac{u(x - h^- e) - u(x)}{h^-}\Big). \tag{2.38}$$

Note that this construction coincides with Definition 2.1.4 when x is sufficiently far from $\partial\Omega$. For smooth u, one has $\delta_h^e u(x) = \langle \nabla u(x), e \rangle + \mathcal{O}(h^r)$ and $\Delta_h^e u(x) = \mathcal{O}(h^r)$ where $r = 1$ if x is close to $\partial\Omega_h$, and $r = 2$ otherwise. In addition the discrete operator defined by

$$-L_h u(x) := \lambda\, \delta_h^e u(x) + \Delta_h^e u(x)$$

is DDE provided $h\lambda \leq 2$, similarly to the constant coefficient case, since $0 < h_x^{\pm e} \leq h$. Therefore (2.37) and (2.38) can be used as a drop-in replacement for the finite difference operators of Definition 2.1.4 when Dirichlet boundary conditions are used, the resulting scheme is DDE under the same conditions. More complex boundary conditions may require ad-hoc treatment.

Semi-linear operators

Let $D \in S_d^{++}$, $d \in \{2, 3\}$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Consider the semi-linear operator $\mathcal{L}$ and its discretization L_h defined by

$$-\mathcal{L}u(x) := g(\nabla u(x)) + \tfrac{1}{2}\operatorname{Tr}(D\nabla^2 u(x)), \qquad -L_h u(x) := g(D^{-1}\nabla_h^D u(x)) + \tfrac{1}{2}\Delta_h^D u(x).$$

The operator $\mathcal{L}$ is degenerate elliptic, since in the continuous setting this property is independent of the first order term of the PDE. On the other hand, the scheme L_h is DDE provided the linear scheme $\tilde{L}_h$ defined by $-\tilde{L}_h u(x) := \langle D^{-1}\omega, \nabla_h^D u(x)\rangle + \frac{1}{2}\Delta_h^D u(x)$ is DDE for all $\omega \in \nabla g(\mathbb{R}^d) = \{\nabla g(x); x \in \mathbb{R}^d\}$. (This is a severe restriction if g is e.g. a quadratic function, but for such applications it can be enough to check that the scheme is DDE in a neighborhood of the solution.)

Fully non-linear operators

Fully non-linear Hamilton-Jacobi-Bellman (HJB) operators can be expressed, under mild regularity assumptions, in Isaacs form

$$\mathcal{L}u(x) := \sup_{\alpha\in A}\inf_{\beta\in B}\mathcal{L}_{\alpha\beta}u(x), \tag{2.39}$$

where A, B are known as the control sets. In addition $\mathcal{L}_{\alpha\beta}$ is a linear degenerate elliptic operator, for all $\alpha \in A$, $\beta \in B$,

$$\mathcal{L}_{\alpha\beta}u(x) := \mu_{\alpha\beta}(x) + \lambda_{\alpha\beta}(x)u(x) + \langle\omega_{\alpha\beta}(x), \nabla u(x)\rangle - \tfrac{1}{2}\operatorname{Tr}(D_{\alpha\beta}(x)\nabla^2 u(x)),$$

where $\mu_{\alpha\beta}(x) \in \mathbb{R}$, $\lambda_{\alpha\beta}(x) \geq 0$, $\omega_{\alpha\beta}(x) \in \mathbb{R}^d$, and $D_{\alpha\beta}(x) \in S_d^+$. In the special case where one of the sets A or B is a singleton, which is common—consider the Monge-Ampère [BCM16] or Pucci (Chapter 5) equations—then (2.39) is known as the Bellman form of the operator.

It is in principle possible to introduce samples $A_h \subset A$ and $B_h \subset B$ of the control sets, and to construct a discretization $L^h_{\alpha\beta}$ of each linear operator $\mathcal{L}_{\alpha\beta}$ following the approach presented in this paper. This produces a DDE approximation of the operator $\mathcal{L}$

$$L_h u(x) := \sup_{\alpha\in A_h}\inf_{\beta\in B_h} L^h_{\alpha\beta}u(x).$$

Let us acknowledge, however, that this construction is far from straightforward to put in practice, especially if the sets A and B are non-compact, and if the condition number of the matrices $D_{\alpha\beta}(x)$ is not uniformly bounded.

2.B Terminology and elementary properties of polyhedra

In this section, we recall some of the terminology and elementary properties related with polyhedra, limiting our attention to those which are immediately useful in the study of Ryskov's polyhedron and its variant §2.2.2 and §2.3.1. See [BG15] for a more complete reference.

2.B.1 Regularity and skeleton

Definition 2.B.1. A *polyhedron* in $\mathbb{R}^n$ is a set of the form

$$\mathcal{M} := \{x \in \mathbb{R}^n;\, \forall i \in I,\, \langle l_i, x\rangle \geq \alpha_i\}, \tag{2.40}$$

where $l_i \in \mathbb{R}^n$, $\alpha_i \in \mathbb{R}$, and I is a finite or countable set. The polyhedron $\mathcal{M}$ is said *regular* iff it (i) has a non-empty interior, (ii) does not contain any affine line, and (iii) can be locally described by the constraints corresponding to a finite subset of I.

By definition, a polyhedron is thus a convex set. Condition (ii) can be reformulated as $\mathrm{Span}\{l_i; i \in I\} = \mathbb{R}^n$. Condition (iii) can be reformulated as follows: for all $x \in \mathcal{M}$ there exists a positive radius $r > 0$ and a finite subset $I_0 \subset I$ such that

$$\langle l_i, y\rangle > \alpha_i, \ \forall i \in I \setminus I_0, \ \forall y \in B(x, r).$$

Definition 2.B.2. Let $\mathcal{M}$ be a regular polyhedron, defined as in (2.40). A k-facet of $\mathcal{M}$, where $1 \leq k \leq n$, is a *non-empty* subset of $\mathcal{M}$ of the form

$$\{x \in \mathcal{M};\ \forall i \in J,\ \langle l_i, x\rangle = \alpha_i\}, \qquad \text{where} \quad \dim \mathrm{Span}\{l_i;\ i \in J\} = n - k,$$

and where $J \subset I$ denotes a subset of the constraint indices.

By construction, a k-facet is a convex subset of $\partial\mathcal{M}$ of affine dimension k. If a k-facet satisfies $\#(J) > n - k$, where $J \subset I$ is chosen maximal for inclusion, then it is said degenerate. By construction 0-facets are singletons, and their single point is called a *vertex*. On the other hand 1-facets are known as *edges* and come in two flavors

- *Bounded edges*, of the form $[x_1, x_2] := \{(1-t)x_1 + tx_2;\ 0 \leq t \leq 1\}$, where x_1 and x_2 are vertices.
- *Unbounded edges*, of the form $\{x + \lambda v;\ \lambda \geq 0\}$, where x is a vertex, and $v \in \mathbb{R}^n \setminus \{0\}$ is called the unbounded edge direction (unique up to multiplication by a positive constant).

Note that *doubly unbounded edges*, of the form $\{x + \lambda v;\ \lambda \in \mathbb{R}\}$, are affine lines and are thus excluded by Definition 2.B.1.

Remark 2.B.3. Let $\mathcal{M}$ be a regular polyhedron, in the sense of Definition 2.B.1. An element $x \in \mathcal{M}$ is a vertex iff $\mathbb{R}^n = \mathrm{Span}\{l_i;\ i \in I, \langle l_i, x\rangle = \alpha_i\}$.

2.B.2 Linear programs

Linear programs are defined as the optimization of a linear functional over a polytope. A fundamental result of operational research, is that such problems can under suitable assumptions be solved by a greedy search over the graph defined by the edges of the polytope, such as the simplex algorithm [BG15]. Since Definition 2.B.1 allows for infinitely many constraints, which is slightly more general than the common setting, we establish in Proposition 2.B.4 a basic result on such programs, used in §2.3.2. Note that the infima in (2.41) may not be attained.

Proposition 2.B.4. *Let $\mathcal{M}$ be a regular polyhedron. Then for any $l \in \mathbb{R}^n$*

$$\begin{aligned} &\inf\{\langle l, x\rangle;\ x \in \mathcal{M}\} \qquad (2.41)\\ &= \begin{cases} -\infty \text{ if } \langle l, v\rangle < 0 \text{ for some unbounded edge direction } v, \\ \inf\{\langle l, x\rangle;\ x \text{ vertex of } \mathcal{M}\} \text{ otherwise.} \end{cases} \end{aligned}$$

Proof. By point (i) of Definition 2.B.1, there exists $x_* \in \mathrm{int}(\mathcal{M})$. By point (ii) of Definition 2.B.1, one has $\mathrm{Span}\{l_i; i \in I\} = \mathbb{R}^n$, otherwise $x_* + \mathbb{R}v$ is an affine line contained in $\mathcal{M}$ for any non-zero $v \in \mathrm{Span}\{l_i; i \in I\}^\perp$, hence there exists $I_* \subset I$ with $\#(I_*) = n$ and such that $(l_i)_{i \in I_*}$ is a basis of $\mathbb{R}^n$.

Define $l_* := \sum_{i \in I_*} l_i$, and consider for each $\alpha > l_*(x_*)$ the set $\mathcal{M}_\alpha := \{x \in \mathcal{M};\ \langle l_*, x\rangle \leq \alpha\}$. Note that for each $x \in \mathcal{M}_\alpha$ and $i \in I_*$ one has $0 \leq l_i(x) - \alpha_i \leq \alpha - \sum_{i \in I_*} \alpha_i$, hence $\mathcal{M}_\alpha$ is bounded. Thus $\mathcal{M}_\alpha$ is a compact polyhedron with non-empty interior, which by Definition

2.B.1 (iii) is characterized by finitely many linear constraints. By Carathéodory's theorem, $\min\{\langle l, x\rangle;\, x \in \mathcal{M}_\alpha\}$ is attained at a vertex of $\mathcal{M}_\alpha$, which by construction is either a vertex of $\mathcal{M}$ or the intersection of an edge of $\mathcal{M}$ (bounded or not) with the hyperplane $\{x \in \mathbb{R}^n;\, \langle l_*, x\rangle = \alpha\}$. From this point, and noting that $\mathcal{M} = \cup_{\alpha\in\mathbb{R}}\mathcal{M}_\alpha$, the announced result easily follows. □

Definition 2.B.5 (Karush-Kuhn-Tucker relations)**.** A set of KKT relations for l in $\mathbb{R}^n$ and x in $\mathcal{M}$ is a finitely supported family of non-negative coefficients $(\lambda_i)_{i\in I}$ such that

$$l = \sum_{i\in I} \lambda_i l_i, \qquad \text{and} \quad \forall i \in I,\ \lambda_i = 0 \ \text{ or } \ \langle l_i, x\rangle = \alpha_i.$$

It is known [BG15] that a linear form $l \in \mathbb{R}^n$ attains its minimum at a given point x of a regular polyhedron $\mathcal{M}$, *if and only if* there exists KKT relations for l and x. The next result establishes a uniqueness property of the KKT relations.

Proposition 2.B.6. *Let $\mathcal{M} \subset \mathbb{R}^n$ be a regular polyhedron, in the sense of Definition 2.B.1. Assume that one has a set of KKT relations $(\lambda_i)_{i\in I}$ for some $l \in \mathbb{R}^d$ at a* non-degenerate *vertex $x \in \mathcal{M}$. Then any other KKT relations $(\lambda'_i)_{i\in I}$ at some $x' \in \mathcal{M}$ (possibly distinct from x), for the same l, obey $\lambda_i = \lambda'_i$ for all $i \in I$.*

Proof. For all $i \in I$ such that $\lambda'_i > 0$ one has $\langle l_i, x'\rangle = \alpha_i$, thus $\langle l_i, x - x'\rangle \geq 0$. On the other hand one has $\langle l, x\rangle = \langle l, x'\rangle = \inf\{\langle l, z\rangle;\, z \in \mathcal{M}\}$, and therefore $0 = \langle l, x - x'\rangle = \sum_{i\in I} \lambda'_i \langle l_i, x - x'\rangle$. Combining these two arguments we obtain that for all $i \in I$ such that $\lambda'_i > 0$ one has $\langle l_i, x - x'\rangle = 0$, and therefore $\langle l_i, x\rangle = \alpha_i$. Since x is a non-degenerate vertex, the family $\{l_i;\, i \in I, \langle l_i, x\rangle = \alpha_i\}$ is a basis of $\mathbb{R}^n$, which implies the announced uniqueness result. □

2.B.3 Edges originating from a vertex

In this section, we present a constructive enumeration of all the edges of a regular polyhedron $\mathcal{M}$ containing a given vertex x. This description follows from Definition 2.B.2 of k-facets, here with $k = 1$. We use the notations of Definition 2.B.1.

Let $J := \{i \in I;\, \langle l_i, x\rangle = \alpha_i\}$ denote the indices of all the active constraints at the vertex x of $\mathcal{M}$. In order to enumerate all the edges of $\mathcal{M}$ containing x, bounded or unbounded, the steps are the following:

(A) Consider successively all subsets S of J with cardinality $n - 1$.

(B) If $\dim \mathrm{Span}\{l_i;\, i \in S\} < n - 1$, then skip this subset. Otherwise denote by $\nu \in \mathbb{R}^n \setminus \{0\}$ the vector, which is unique up to a scalar multiplication, such that $\langle l_i, \nu\rangle = 0$ for all $i \in S$.

(C) Replace ν with its opposite $-\nu$, if necessary, in such way that $\langle l_i, \nu\rangle \geq 0$ for all $i \in J \setminus S$. If that is not possible, then skip this subset.

(D) Compute $\Lambda := \sup\{\lambda \in \mathbb{R}; x + \lambda\nu \in \mathcal{M}\}$. If $\Lambda = +\infty$, then there is an unbounded edge at x in the direction of ν. Otherwise, x and $x + \Lambda\nu$ are the vertices of a bounded edge of $\mathcal{M}$.

Chapter 3

Monotone discretization of anisotropic four-dimensional differential operators using Voronoi's first reduction

3.1 Introduction

In this chapter, we develop monotone finite differences for anisotropic diffusion operators. Our approach requires a Cartesian grid discretization, and is applicable in dimension $d \leq 4$. It leverages tools from Euclidean lattice geometry, and more specifically Voronoi's first reduction of quadratic forms [Sch09a].

Let us denote by $\mathcal{S}_d$ (respectively $\mathcal{S}_d^+$, $\mathcal{S}_d^{++}$) the set of symmetric (respectively symmetric positive semidefinite, symmetric positive definite) matrices of size d. The anisotropic diffusion operators that we consider are of the form

$$u \mapsto \mathrm{Tr}(\mathcal{D}(\cdot)D^2u(\cdot)),$$

where $u \in C^2(\mathbb{R}^d)$, and $\mathcal{D}\colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ is a given field of positive definite matrices. We develop adaptive finite difference discretizations of those operators on Cartesian grids, taking into account the preferred directions defined by matrices of the field $\mathcal{D}$, which are typically anisotropic and whose eigenvectors are not aligned with the discretization grid. For that purpose, we need to introduce some notation: let $\mathcal{Z}_d$ refer to the collection of nonzero vectors of size d with integer entries, keeping only one representative among pairs of opposites, and let Λ_d collect all finitely supported and nonnegative maps on $\mathcal{Z}_d$:

$$\mathcal{Z}_d := (\mathbb{Z}^d \setminus \{0\})/\pm, \qquad \Lambda_d := \{\lambda = (\lambda^e)_{e\in\mathcal{Z}_d}\colon \mathcal{Z}_d \to \mathbb{R}_+, \text{ finitely supported}\}.$$

For any $h > 0$, $e \in \mathcal{Z}_d$, $u\colon \mathbb{R}^d \to \mathbb{R}$, and $x \in \mathbb{R}^d$, we introduce the second-order finite difference

$$\Delta_h^e u(x) := \frac{u(x+he) + u(x-he) - 2u(x)}{h^2}.$$

In this chapter we construct, and study, coefficients $\lambda(x) \in \Lambda_d$ such that the following approxima-

tion holds for sufficiently smooth u:

$$\mathrm{Tr}(\mathcal{D}(x)D^2u(x)) \approx \sum_{e\in\mathcal{Z}_d} \lambda^e(x)\Delta_h^e u(x). \tag{3.1}$$

Definition 3.1.1. A family of coefficients $\lambda\colon \mathbb{R}^d \to \Lambda_d$, denoted $\lambda = (\lambda^e(x))_{x\in\mathbb{R}^d}^{e\in\mathcal{Z}_d}$ is said:

- $\mathcal{D}$*-consistent*, where $\mathcal{D}\colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ is a field of positive definite matrices, if, for any $x \in \mathbb{R}^d$, one has
$$\mathcal{D}(x) = \sum_{e\in\mathcal{Z}_d} \lambda^e(x) ee^\top. \tag{3.2}$$
- K*-Lipschitz* if, for any x, $y \in \mathbb{R}^d$ and $e \in \mathcal{Z}_d$, one has
$$|\lambda^e(x) - \lambda^e(y)| \leq K|x - y|.$$
- R*-supported* if $r(x) \leq R$ for any $x \in \mathbb{R}^d$, where
$$r(x) := \max\{|e| \mid e \in \mathcal{Z}_d,\ \lambda^e(x) > 0\}.$$
- ε*-spanning* if, for any $x \in \mathbb{R}^d$, one has
$$\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e(x) \geq \varepsilon\} = \mathbb{Z}^d,$$
where $\mathrm{Span}_{\mathbb{Z}}(E)$ denote the collection of all linear combinations with integer coefficients of elements of a set $E \subset \mathbb{Z}^d$.

We propose in this chapter a practical and efficient method for constructing coefficients λ complying with above properties, see Corollary 3.1.4. In the following points, we discuss Definition 3.1.1 and contrast our approach with two-scale discretizations of PDEs, see [DJ13; NNZ19].

- $\mathcal{D}$-consistency is a qualitative property, ensuring that (3.1) is second order accurate with respect to the grid scale $h > 0$. While we choose this definition, it would also make sense to consider a quantitative variant, featuring a consistency error. For instance two-scale discretizations [DJ13; NNZ19] feature such an error, depending on an intermediate scale satisfying $h \ll k \ll 1$, and vanishing $k \to 0$. A consistency error is also unavoidable if one addresses rank deficient semi-definite diffusion matrices [MW52], unless their kernel is spanned by vectors with integer entries.
- To the knowledge of the authors, all practical finite difference schemes involving adaptive matrix decompositions (3.2) feature coefficients with Lipschitz regularity, but *not better*. See [DJ13; BOZ04; FM14; Wei98], which include two-scale methods. Lipschitz regularity is sometimes sufficient to establish convergence rates [BJ07], although in some cases better rates could be obtained if the coefficients satisfied stronger regularity assumptions, see for instance [Kry05] where the square roots of the coefficients have to be Lipschitz continuous.
- R-support is a quantitative property, controlling the effective scale $k = Rh$ of the numerical scheme. The radius R is bounded, for our numerical scheme, in terms of the maximal condition number of the matrix field $\mathcal{D}$. In contrast, two scale discretizations of PDEs involve an effective discretization scale which decreases sub-linearly with h, for instance $k = h^{2/5}$ is optimal in [DJ13], which yields reduced convergence rates.
- The ε-spanning assumption ensures that the graph underlying the diffusion is locally connected. This property guarantees that the numerical solution does not suffer from checkerboard artifacts, see Proposition 3.5.5 and Corollary 3.5.6.

Voronoi's decomposition of matrices. When discretizing anisotropic diffusion with adaptive finite differences, the scheme coefficients should obey a number of properties summarized in Definition 3.1.1. In the following, we describe an efficient method for computing such coefficients, leveraging a tool from discrete geometry known as Voronoi's first reduction of quadratic forms [Sch09a]. Our numerical scheme is obtained by solving an optimization problem, for each diffusion matrix $\mathcal{D} = \mathcal{D}(x)$, where x is a (discretization) point of the PDE domain. More precisely, we solve for $\mathcal{D} \in \mathcal{S}_d^{++}$:

$$\Lambda(\mathcal{D}) := \operatorname*{argmax}_{\lambda \in \Lambda_d} \left\{ \sum_{e \in \mathcal{Z}_d} \lambda^e \;\middle|\; \sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top = \mathcal{D} \right\}. \tag{3.3}$$

We call (3.3) a linear program, although it is posed in an infinite dimensional space; see Proposition 3.2.3 for further discussion. This linear program benefits from multiple invariances and symmetries, related to linear changes of coordinates with integer coefficients, and for this reason it can be solved extremely efficiently.

One can show that $\Lambda(\mathcal{D})$ is a singleton in dimension $d \leq 3$ and either a singleton or an equilateral triangle if $d = 4$, see section 3.3. In the latter two cases, one must thus select an element from this set, which motivates the following definition.

Definition 3.1.2. For any $\mathcal{D} \in \mathcal{S}_d^{++}$, we define $\lambda(\mathcal{D}) \in \Lambda_d$ as follows:

- If $d \in \{1, 2, 3\}$, then $\lambda(\mathcal{D})$ is defined as the unique element of $\Lambda(\mathcal{D})$.
- If $d = 4$, then $\lambda(\mathcal{D})$ is defined as the barycenter of $\Lambda(\mathcal{D})$.

For any $\mathcal{D} \in \mathcal{S}_d^{++}$, let us denote by $\mu(\mathcal{D}) \geq 1$ the square root of its condition number:

$$\mu(\mathcal{D}) := \sqrt{|\mathcal{D}||\mathcal{D}^{-1}|}.$$

Theorem 3.1.3. *The mapping $\lambda \colon \mathcal{S}_d^{++} \to \Lambda_d$ of Definition 3.1.2 obeys the following properties, where $\mathcal{D} \in \mathcal{S}_d^{++}$ is arbitrary and where $\mu := \mu(\mathcal{D})$ denotes the square root of its condition number:*

- *It is consistent, in the sense that $\mathcal{D} = \sum_{e \in \mathcal{Z}_d} \lambda^e(\mathcal{D}) e e^\top$.*
- *It is $K(\mu)$-Lipschitz, that is $|\lambda^e(\mathcal{D}_1) - \lambda^e(\mathcal{D}_2)| \leq K(\mu)|D_1 - D_2|$ if $e \in \mathcal{Z}_d$, $\mathcal{D}_1$, $\mathcal{D}_2 \in \mathcal{S}_d^{++}$, and $\max\{\mu(\mathcal{D}_1), \mu(\mathcal{D}_2)\} \leq \mu$.*
- *It is $R(\mu)$-supported, that is $|e| \leq R(\mu)$ for any $e \in \mathcal{Z}_d$ such that $\lambda^e(\mathcal{D}) > 0$.*
- *It is ε-spanning, that is $\operatorname{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e(\mathcal{D}) \geq \varepsilon |\mathcal{D}^{-1}|^{-1}\} = \mathbb{Z}^d$, where $\varepsilon > 0$ depends only on d.*

Furthermore, one has $K(\mu) = O(\mu^2)$ and $R(\mu) = O(\mu)$.

Corollary 3.1.4. *Let $\mathcal{D} \colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ be K-Lipschitz and have bounded condition number and uniformly positive smallest eigenvalue. Define $\lambda^e(x) := \lambda^e(\mathcal{D}(x))$, for any $x \in \mathbb{R}^d$ and $e \in \mathcal{Z}_d$, where λ is from Definition 3.1.2. Then $\lambda \colon \mathbb{R}^d \to \Lambda_d$ is $\mathcal{D}$-consistent, K'-Lipschitz, R-supported, and ε-spanning. The constants K', R, and $\varepsilon > 0$ only depend on d, K, $\|\mu(\mathcal{D})\|_\infty$, and $\|\mathcal{D}^{-1}\|_\infty$.*

We emphasize that Definition 3.1.2 is completely practical, in the sense that $\lambda(\mathcal{D})$ can be computed in a fast and reliable manner numerically, see section 3.3.

The consistency property in Theorem 3.1.3 follows from the definition of λ. The $K(\mu)$-Lipschitz regularity, $R(\mu)$-supportedness, and ε-spanning properties are proved respectively in Theorem 3.3.6, Theorem 3.4.1, and Theorem 3.5.1. Corollary 3.1.4 follows immediately from Theorem 3.1.3.

3.2 Voronoi's first reduction of quadratic forms

Voronoi's first reduction [Vor08] is a tool from the field of lattice geometry [Sch09a], with applications in sphere packing, arithmetic, and PDE discretizations in this chapter and [Mir19]. It is originally intended for classifying positive quadratic forms up linear changes of coordinates stabilizing the lattice $\mathbb{Z}^d$, represented by the set $\mathrm{GL}_d(\mathbb{Z})$ of matrices with integer entries $A \in \mathbb{Z}^{d\times d}$ satisfying $\det(A) = \pm 1$.

Definition 3.2.1 (Arithmetical equivalence)**.** Two matrices $M_1,\ M_2 \in \mathcal{S}_d$ are arithmetically equivalent if there exists $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $M_2 = A^\top M_1 A$.

Voronoi's first reduction $\mathrm{Vor}(\mathcal{D})$ of $\mathcal{D} \in \mathcal{S}_d^{++}$ is defined similarly to a linear program, although with infinitely many constraints. Its modern presentation involves an auxiliary object $\mathcal{M}_d \subset \mathcal{S}_d$, referred to as Ryskov's polyhedron:

$$\mathrm{Vor}(\mathcal{D}) := \min_{M\in\mathcal{M}_d} \mathrm{Tr}(\mathcal{D}M), \qquad \mathcal{M}_d := \{M \in \mathcal{S}_d \mid \forall e \in \mathcal{Z}_d,\ \langle e, Me\rangle \geq 1\}. \tag{3.4}$$

This optimization problem is well-posed, as proved by Voronoi himself [Vor08; Sch09a].

Theorem 3.2.2 (Voronoi)**.** *Ryskov's polyhedron is a subset of $\mathcal{S}_d^{++}$ and is a locally finite polyhedron, in the sense that finitely many constraints are actively locally in the neighborhood of any point. It has finitely many equivalence classes of vertices for the relation of arithmetical equivalence. The linear program* $\mathrm{Vor}(\mathcal{D})$ *is well-posed in the sense that the collection of minimizers is non-empty and compact.*

For any $M \in \mathcal{S}_d^{++}$, we define $\Xi(M) := \{e \in \mathcal{Z}_d \mid \langle e, Me\rangle \leq 1\}$. If $M \in \mathcal{M}_d$, then $\Xi(M)$ denotes the set of active constraints in (3.4), and $\Xi(M)$ is finite by Theorem 3.2.2.

We establish below duality relations between the linear program (3.3) defining our discretization and Voronoi's first reduction (3.4).

Proposition 3.2.3. *Let $\mathcal{D} \in \mathcal{S}_d^{++}$, and let $M \in \mathcal{M}_d$ be optimal in* (3.4)*. Then the set $\Lambda(\mathcal{D})$ of maximizers in* (3.3) *is a nonempty convex compact polytope characterized by*

$$\Lambda(\mathcal{D}) = \left\{ \lambda \in \Lambda_d \;\middle|\; \lambda^e > 0 \implies e \in \Xi(M),\ \sum_{e\in\Xi(M)} \lambda^e ee^\top = \mathcal{D} \right\}. \tag{3.5}$$

Proof. For now, we waive the constraint that λ is finitely supported in (3.3), and we define

$$\Lambda'(\mathcal{D}) := \operatorname*{argmax}_{\lambda\in l^1_w(\mathcal{Z}_d)} \left\{ \sum_{e\in\mathcal{Z}_d} \lambda^e \;\middle|\; \lambda \succeq 0,\ \sum_{e\in\mathcal{Z}_d} \lambda^e ee^\top = \mathcal{D} \right\} = \operatorname*{argmax}_{\lambda\in l^1_w(\mathcal{Z}_d)} (-f(\lambda) - g(A\lambda)), \tag{3.6}$$

where the vector space $l^1_w(\mathcal{Z}_d) := \{\lambda\colon \mathcal{Z}_d \to \mathbb{R} \mid |\lambda|_{l^1_w} < +\infty\}$ is equipped with the norm $|\cdot|_{l^1_w} : \lambda \mapsto \sum_{e\in\mathcal{Z}_d} |e|^2|\lambda^e|$, and where

$$f(\lambda) := \chi_{\{\lambda\succeq 0\}} - \sum_{e\in\mathcal{Z}_d} \lambda^e, \qquad A\lambda := \sum_{e\in\mathcal{Z}_d} \lambda^e ee^\top, \qquad g(P) := \chi_{\{P=\mathcal{D}\}}.$$

The choice of the norm $|\cdot|_{l^1_w}$ is justified by the fact that any admissible λ in (3.6) satisfies

$$|\lambda|_{l^1_w} = \mathrm{Tr}\left(\sum_{e\in\mathcal{Z}_d} \lambda^e ee^\top\right) = \mathrm{Tr}(\mathcal{D}) < +\infty. \tag{3.7}$$

The minimization problem (3.4) is the dual to (3.6), in the sense of Fenchel's duality theorem. Therefore, if $M \in \mathcal{M}_d$ is optimal in (3.4), then for any $\lambda \in l^1_w$ admissible in (3.6),

$$0 \leq \mathrm{Vor}(\mathcal{D}) - \sum_{e \in \mathcal{Z}_d} \lambda^e = \mathrm{Tr}(\mathcal{D}M) - \sum_{e \in \mathcal{Z}_d} \lambda^e = \sum_{e \in \mathcal{Z}_d} \lambda^e (\langle e, Me \rangle - 1). \tag{3.8}$$

Moreover, the constraint qualification condition $0 \in \mathrm{int}(\mathrm{dom}\, g - A\,\mathrm{dom}\, f)$ (equivalently $\mathcal{D} \in \mathrm{int}(A\,\mathrm{dom}\, f)$, where $A\,\mathrm{dom}\, f = \{\sum_{e \in \mathcal{Z}^d} \lambda^e e e^\top \mid \lambda \in l^1_w(\mathcal{Z}_d),\ \lambda \succeq 0\}$) is satisfied, since $\mathcal{D}$ may be approximated by symmetric positive definite matrices with rational eigenvectors. Therefore the inequality in (3.8) is an equality if and only if $\lambda \in \Lambda'(\mathcal{D})$. Using that all terms in the right-hand side of (3.8) are nonnegative, we deduce that an admissible λ in (3.6) belongs to $\Lambda'(\mathcal{D})$ if and only if it is supported on $\Xi(M)$. In particular, any $\lambda \in \Lambda'(\mathcal{D})$ is finitely supported, thus $\Lambda(\mathcal{D}) = \Lambda'(\mathcal{D})$ and (3.5) holds. The compactness of $\Lambda(\mathcal{D})$ follows from the fact that any λ in the finite-dimensional set $\Lambda(\mathcal{D})$ satisfies (3.7). □

In order to proceed with the proof of Theorem 3.1.3, we need a more precise description of Ryskov's polyhedron. We define the set

$$\mathrm{Perfect}(d) := \{M \in \mathcal{M}_d \mid \mathrm{Span}_{\mathbb{R}}\{e e^\top \mid e \in \Xi(M)\} = \mathcal{S}_d\}$$

of vertices of $\mathcal{M}_d$, which are known as *perfect forms* [Vor08]. For any perfect form $M \in \mathrm{Perfect}(d)$, we define the set

$$\mathcal{N}(M) := \{M' \in \mathrm{Perfect}(d) \mid \dim(\mathrm{Span}_{\mathbb{R}}\{e e^\top \mid e \in \Xi(M) \cap \Xi(M')\}) = d(d+1)/2 - 1\},$$

of neighbor vertices of M in $\mathcal{M}_d$, where $d(d+1)/2 = \dim(\mathcal{S}_d)$. The polyhedral structure of $\mathcal{M}_d$ is compatible with the relation of arithmetical equivalence defined in Definition 3.2.1:

Proposition 3.2.4. *If $M \in \mathrm{Perfect}(d)$ and $A \in \mathrm{GL}_d(\mathbb{Z})$, then $A^\top M A \in \mathrm{Perfect}(d)$ and*

$$\Xi(A^\top M A) = \{A^{-1} e \mid e \in \Xi(M)\}, \qquad \mathcal{N}(A^\top M A) = \{A^\top M' A \mid M' \in \mathcal{N}(M)\}.$$

Proof. This follows directly from the definitions of $\mathcal{M}_d$, $\Xi(M)$, and $\mathcal{N}(M)$, and from the fact that for any $A \in \mathrm{GL}_d(\mathbb{Z})$ one has $\{Ae \mid e \in \mathcal{Z}_d\} = \mathcal{Z}_d$. □

The classification of perfect forms up to arithmetical equivalence is a classical problem in lattice geometry [CS88], whose complexity explodes as dimension increases, see [DSV07] for the latest complete classification in dimension $d = 8$. Fortunately, we are only interested in $d \leq 4$. There is a canonical perfect form, existing in arbitrary dimension d, and defined as follows: denoting $\mathbb{1} := (1, \ldots, 1) \in \mathbb{Z}^d$,

$$M^*_d := \frac{1}{2}(I_d + \mathbb{1}\mathbb{1}^\top) = \frac{1}{2} \begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \ldots & 1 & 2 \end{pmatrix}.$$

The next proposition confirms that $\mathrm{Span}_{\mathbb{R}}\{e e^\top \mid e \in \Xi(M^*_d)\} = \mathcal{S}_d$ and thus that M^*_d is a perfect form.

Proposition 3.2.5. *For any $e \in \mathcal{Z}_d$, one has $\langle e, M^*_d e \rangle \geq 1$, with equality if and only if $e = \pm e_i$, for some $1 \leq i \leq d$, or $e = \pm(e_i - e_j)$, for some $1 \leq i < j \leq d$, where $(e_i)_{1 \leq i \leq d}$ denotes the canonical basis of $\mathbb{R}^d$.*

Proof. Let $e \in \mathcal{Z}_d$ be such that $1 = \langle e, M_d^* e\rangle = (|e|^2 + \langle e, \mathbb{1}\rangle^2)/2$. Then $|e|^2 \leq 2$, and therefore e has either one or two nonzero components, equal to ± 1. In the latter case these components have opposite sign, since $\langle e, \mathbb{1}\rangle^2 = 0$. The result follows. □

In dimension $d = 4$, the following is also a perfect form [CS88] (and is not arithmetically equivalent to M_4^* since it does not have the same determinant):

$$M_4' := \frac{1}{2}\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

Proposition 3.2.6. *The matrix M_4' is a perfect form, and*

$$\begin{aligned}\Xi(M_4') = \{\pm e_i \mid 1 \leq i \leq 4\} \cup \{\pm(e_i - e_j) \mid 1 \leq i < j \leq 4,\, \{i,j\} \neq \{1,4\}\} \\ \cup \{\pm(e_1 - e_i + e_4) \mid 2 \leq i \leq 3\} \cup \{\pm(e_1 - e_2 - e_3 + e_4)\},\end{aligned}$$

where $(e_i)_{1\leq i\leq d}$ denotes the canonical basis of $\mathbb{R}^d$.

Proof. We compute $\Xi(M_4')$ using a computer assisted procedure. By Lemma 3.2.7 below, it suffices to check, for any of the finitely many $e \in \mathcal{Z}_d$ satisfying $|e|^2 \leq \lambda_{\min}(M_4')^{-1}$, whether $e \in \Xi(M_4')$ (note that $\lambda_{\min}(M_4')^{-1} = (5 + \sqrt{17})/2 \approx 4.56$). We also compute that any $e \in \Xi(M_4')$ satisfies the equality $\langle e, Me\rangle = 0$, hence $M_4' \in \mathcal{M}_d$. Since $\mathrm{Span}_{\mathbb{R}}\{ee^\top \mid e \in \Xi(M_4')\} = \mathcal{S}_d$, the matrix M_4' is a perfect form. □

Lemma 3.2.7. *If $M \in \mathcal{S}_d^{++}$ and $e \in \Xi(M)$, then $|e|^2 \leq \lambda_{\min}(M)^{-1}$.*

Proof. One has $\lambda_{\min}(M)|e|^2 \leq \langle e, Me\rangle \leq 1$. □

Comparing the cardinalities of the sets $\Xi(M_4^*)$ and $\Xi(M_4')$ of active constraints at points M_4^* and M_4' (respectively 10 and 12) with the dimension $\dim(\mathcal{S}_4) = 10$ of the optimization space, we find that M_4' is a degenerate vertex of Ryskov's polyhedron $\mathcal{M}_4$, whereas M_4^* is a nondegenerate vertex.

In dimension $d \leq 3$, there is only one equivalence class of perfect forms for the relation of arithmetical equivalence, associated with the representative M_d^*, and for this reason Voronoi's first reduction (3.4) is particularly simple to study and compute, using Selling's algorithm [Sel74; CS92]. In contrast there is in dimension $d = 4$ one additional equivalence class of perfect forms, associated with the representative M_4' [CS88]. In the following, it will be convenient to express those facts in terms of a finite subset $\mathrm{Perfect}_0(d)$ of $\mathrm{Perfect}(d)$ satisfying

$$\mathrm{Perfect}(d) = \{A^\top M A \mid A \in \mathrm{GL}_d(\mathbb{Z}),\, M \in \mathrm{Perfect}_0(d)\}. \tag{3.9}$$

Definition 3.2.8. We let:

- $\mathrm{Perfect}_0(d) := \{M_d^*\}$, if $d \leq 3$.
- $\mathrm{Perfect}_0(4) := \{M_4^*, M_4'\}$.
- $\mathrm{Perfect}_0(d)$ be an arbitrary finite subset of $\mathrm{Perfect}(d)$ satisfying (3.9), if $d \geq 5$.

In dimension $d \geq 5$, the existence of a suitable set $\mathrm{Perfect}_0(d)$ follows from Theorem 3.2.2. The following proposition implies that (3.9) is still satisfied in dimension $d \leq 4$.

Proposition 3.2.9. *Let $d \leq 4$. Then for any perfect form $M \in \mathrm{Perfect}(d)$, there exists $M_0 \in \mathrm{Perfect}_0(d)$ and $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $M = A^\top M_0 A$.*

Proof. By Proposition 3.2.4, it suffices to check, for any $M \in \mathrm{Perfect}_0(d)$, that any $M' \in \mathcal{N}(M)$ is arithmetically equivalent to some $M_0 \in \mathrm{Perfect}_0(d)$. This enumeration technique is known as Voronoi's algorithm [Mar03; Vor08], and has been applied successfully in the dimensions considered here. See also section 3.A for a description of a computer assisted procedure implementing Voronoi's algorithm. □

After applying the procedure discussed in section 3.A, we observe that, in dimension $d = 4$, all 10 neighbors of the nondegenerate vertex M_4^* of $\mathcal{M}_d$ are arithmetically equivalent to M_4', while the degenerate vertex M_4' has 48 neighbors arithmetically equivalent to M_4^* and 16 neighbors arithmetically equivalent to M_4'.

3.3 Computing the decomposition

We explain in this section how one may compute in practice, for any matrix $\mathcal{D} \in \mathcal{S}_d^{++}$, the decomposition $\lambda(\mathcal{D})$ defined in Definition 3.1.2. We recommend solving first the minimization problem (3.4), and then using Proposition 3.2.3 to deduce the value of $\lambda(\mathcal{D})$.

Algorithm 2 Solving Voronoi's first reduction — abstract version

Initialization: Let $M \in \mathrm{Perfect}(d)$ (for instance $M \leftarrow M_d^*$).
While there exists $M' \in \mathcal{N}(M)$ such that $\mathrm{Tr}(\mathcal{D}M') < \mathrm{Tr}(\mathcal{D}M)$ **do** $M \leftarrow M'$.
Return M.

Algorithm 3 Solving Voronoi's first reduction — practical version

Initialization:
  Let $M_0 \in \mathrm{Perfect}_0(d)$ (for instance $M \leftarrow M_d^*$).
  Let $A \in \mathrm{GL}_d(\mathbb{Z})$ (for instance $A \leftarrow I_d$).
While there exist $M_0' \in \mathrm{Perfect}_0(d)$ and $A' \in \mathrm{GL}_d(\mathbb{Z})$
  such that $(A')^\top M_0' A' \in \mathcal{N}(M_0)$ and $\mathrm{Tr}(\mathcal{D}A^\top (A')^\top M_0' A' A) < \mathrm{Tr}(\mathcal{D}A^\top M_0 A)$
do
  $M_0 \leftarrow M_0'$.
  $A \leftarrow A'A$.
Return M_0 and A.

Since the cost minimized in (3.4) is linear, the minimum in attained at some vertex of Ryskov's polyhedron $\mathcal{M}_d$. We recommend solving the problem (3.4) by iterating over perfect forms, in the manner described in Algorithm 2. This algorithm is not directly implementable since we did not explain how the set $\mathcal{N}(M)$ is computed, for an arbitrary perfect form M. In practice, in order to benefit from the symmetries of Ryskov's polyhedron, we represent a perfect form M by a pair (M_0, A), where $M_0 \in \mathrm{Perfect}_0(d)$, $A \in \mathrm{GL}_d(\mathbb{Z})$, and $M = A^\top M_0 A$. This yields Algorithm 3, which is equivalent to Algorithm 2 as shown by Proposition 3.2.4. We still need to know, for any $M_0 \in \mathrm{Perfect}_0(d)$, how to express each element of $\mathcal{N}(M_0)$ in the form $(A')^\top M_0' A'$, where $M_0' \in \mathrm{Perfect}_0(d)$ and $A' \in \mathrm{GL}_d(\mathbb{Z})$. Fortunately, those adjacency relations, which do not depend on the matrix $\mathcal{D}$, may be precomputed using well-known algorithms, see section 3.A.

Proposition 3.3.1. *Algorithm 2 terminates and returns a perfect form M that is a minimizer in* (3.4)*. Equivalently, Algorithm 3 terminates and returns a perfect form $M_0 \in \mathrm{Perfect}_0(d)$ and a matrix $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $A^\top M_0 A$ is a minimizer in* (3.4)*.*

Proof. By Theorem 3.2.2, for any $\alpha > 0$, the set $\{M \in \mathcal{M}_d \mid \mathrm{Tr}(\mathcal{D}M) \leq \alpha\}$ is a bounded polyhedron, in particular there are finitely many perfect forms M such that $\mathrm{Tr}(\mathcal{D}M) \leq \alpha$. Thus Algorithm 2 iterates over finitely many perfect forms M. Since the cost $\mathrm{Tr}(\mathcal{D}M)$ decreases strictly at each iteration, the algorithm terminates. The returned M satisfies $\mathrm{Tr}(\mathcal{D}M) \leq \mathrm{Tr}(\mathcal{D}M')$ for any $M' \in \mathcal{N}(M)$, therefore it is a minimizer in (3.4). □

Algorithm 3 returns a decomposition of the minimizer in the form $A^\top M_0 A$ where $M_0 \in \mathrm{Perfect}_0(d)$ and $A \in \mathrm{GL}_d(\mathbb{Z})$. This is useful since then, by the following proposition, we only need to know how to compute $\lambda(\mathcal{D})$ for matrices $\mathcal{D} \in \mathcal{S}_d^{++}$ such that some $M_0 \in \mathrm{Perfect}_0(d)$ is optimal in (3.4).

Proposition 3.3.2. *Let $\mathcal{D} \in \mathcal{S}_d^{++}$ and let $M_0 \in \mathrm{Perfect}_0(d)$ and $A \in \mathrm{GL}_d(\mathbb{Z})$ be such that $A^\top M_0 A$ is a minimizer in* (3.4)*. Then M_0 is a minimizer in* (3.4) *after replacing $\mathcal{D}$ by $A\mathcal{D}A^\top$ (that is, $\mathrm{Vor}(A\mathcal{D}A^\top) = \mathrm{Tr}(A\mathcal{D}A^\top M_0)$), and*

$$\Lambda(\mathcal{D}) = \{(\lambda^{Ae})_{e\in\mathcal{Z}_d} \mid \lambda \in \Lambda(A\mathcal{D}A^\top)\}, \qquad \lambda(\mathcal{D}) = (\lambda^{Ae}(A\mathcal{D}A^\top))_{e\in\mathcal{Z}_d}. \tag{3.10}$$

Proof. We deduce from the equality $\mathcal{M}_d = \{A^\top M A \mid M \in \mathcal{M}_d\}$ that

$$\begin{aligned}\mathrm{Vor}(A\mathcal{D}A^\top) &= \min_{M\in\mathcal{M}_d} \mathrm{Tr}(A\mathcal{D}A^\top M) = \min_{M\in\mathcal{M}_d} \mathrm{Tr}(\mathcal{D}A^\top M A) = \min_{M\in\mathcal{M}_d} \mathrm{Tr}(\mathcal{D}M) = \mathrm{Vor}(\mathcal{D}) \\ &= \mathrm{Tr}(\mathcal{D}A^\top M_0 A) = \mathrm{Tr}(A\mathcal{D}A^\top M_0).\end{aligned}$$

The equalities (3.10) follow directly from Proposition 3.2.3, Proposition 3.2.4, and the fact that $A \in \mathrm{GL}_d(\mathbb{Z})$. □

We describe below how to compute $\lambda(\mathcal{D})$ when some $M_0 \in \mathrm{Perfect}_0(d)$ is optimal in (3.4), in dimension $d \leq 4$.

Proposition 3.3.3. *Let $\mathcal{D} \in \mathcal{S}_d^{++}$, and assume that M_d^* is a minimizer in* (3.4)*. Let $(e_i)_{1\leq i\leq d}$ denote the canonical basis of $\mathbb{R}^d$ and let $\mathcal{D}_{ij}$ denote the component of the matrix $\mathcal{D}$ with indices i and j. Then $\Lambda(\mathcal{D})$ is a singleton, and*

$$\lambda^e(\mathcal{D}) = \begin{cases} \sum_{j=1}^d \mathcal{D}_{ij} & \text{if } e = \pm e_i,\ 1 \leq i \leq d, \\ -\mathcal{D}_{ij} & \text{if } e = \pm(e_i - e_j),\ 1 \leq i < j \leq d, \\ 0 & \text{else.} \end{cases} \tag{3.11}$$

Proof. By Proposition 3.2.3, if $\lambda \in \Lambda(\mathcal{D})$, then $\lambda^e = 0$ for any $e \notin \Xi(M_d^*)$, where $\Xi(M_d^*)$ is described in Proposition 3.2.5, and moreover $\sum_{e\in\Xi(M)} \lambda^e e e^\top = \mathcal{D}$. Since $\#(\{ee^\top \mid e \in \Xi(M_d^*)\}) = d(d+1)/2 = \dim(\mathcal{S}_d)$ and $\mathrm{Span}_{\mathbb{R}}\{ee^\top \mid e \in \Xi(M_d^*)\} = \mathcal{S}_d$, there exists exactly one $\lambda \in \Lambda^d$ satisfying the above properties, and it suffices to check that this is the one defined by (3.11). □

Proposition 3.3.4. *Let $\mathcal{D} \in \mathcal{S}_4^{++}$ and assume that M_4' is a minimizer in* (3.4)*. Let $(e_i)_{1\leq i\leq 4}$ denote the canonical basis of $\mathbb{R}^4$ and let $\mathcal{D}_{ij}$ denote the component of the matrix $\mathcal{D}$ with indices i*

and j. For any α, β, $\gamma \in \mathbb{R}$, let $\lambda_{\alpha,\beta,\gamma}(\mathcal{D}) \in \Lambda_4$ be defined by

$$\lambda^e_{\alpha,\beta,\gamma}(\mathcal{D}) := \begin{cases} \mathcal{D}_{i1} + \mathcal{D}_{i2} + \mathcal{D}_{i3} + \mathcal{D}_{i4} + \gamma & \text{if } e = \pm e_i,\ i \in \{1,4\}, \\ \mathcal{D}_{21} + \mathcal{D}_{22} + \mathcal{D}_{23} + \mathcal{D}_{24} + \alpha & \text{if } e = \pm e_2, \\ \mathcal{D}_{31} + \mathcal{D}_{32} + \mathcal{D}_{33} + \mathcal{D}_{34} + \beta & \text{if } e = \pm e_3, \\ -\mathcal{D}_{i2} - \mathcal{D}_{14} + \beta & \text{if } e = \pm(e_i - e_2),\ i \in \{1,4\}, \\ -\mathcal{D}_{i3} - \mathcal{D}_{14} + \alpha & \text{if } e = \pm(e_i - e_3),\ i \in \{1,4\}, \\ -\mathcal{D}_{23} + \mathcal{D}_{14} + \gamma & \text{if } e = \pm(e_2 - e_3), \\ \alpha & \text{if } e = \pm(e_1 - e_2 + e_4), \\ \beta & \text{if } e = \pm(e_1 - e_3 + e_4), \\ \mathcal{D}_{14} + \gamma & \text{if } e = \pm(e_1 - e_2 - e_3 + e_4), \\ 0 & \text{else.} \end{cases}$$

Then $\Lambda(\mathcal{D})$ is the equilateral triangle characterized by

$$\Lambda(\mathcal{D}) = \{\lambda_{\alpha,\beta,\gamma}(\mathcal{D}) \mid \alpha \geq \alpha_*(\mathcal{D}),\ \beta \geq \beta_*(\mathcal{D}),\ \gamma \geq \gamma_*(\mathcal{D}),\ \alpha + \beta + \gamma = 0\}$$

where

$$\begin{aligned} \alpha_*(\mathcal{D}) &:= \max\{-\mathcal{D}_{21} - \mathcal{D}_{22} - \mathcal{D}_{23} - \mathcal{D}_{24},\ \mathcal{D}_{13} + \mathcal{D}_{14},\ \mathcal{D}_{34} + \mathcal{D}_{14},\ 0\}, \\ \beta_*(\mathcal{D}) &:= \max\{-\mathcal{D}_{31} - \mathcal{D}_{32} - \mathcal{D}_{33} - \mathcal{D}_{34},\ \mathcal{D}_{12} + \mathcal{D}_{14},\ \mathcal{D}_{24} + \mathcal{D}_{14},\ 0\}, \\ \gamma_*(\mathcal{D}) &:= \max\{-\mathcal{D}_{11} - \mathcal{D}_{12} - \mathcal{D}_{13} - \mathcal{D}_{14},\ -\mathcal{D}_{41} - \mathcal{D}_{42} - \mathcal{D}_{43} - \mathcal{D}_{44},\ \mathcal{D}_{23} - \mathcal{D}_{14},\ -\mathcal{D}_{14}\}. \end{aligned}$$

Proof. Let $\Lambda_*(\mathcal{D}) := \{\lambda\colon \Xi(M_4') \to \mathbb{R}_+ \mid \lambda^e \neq 0 \implies e \in \Xi(M_4'),\ \sum_{e \in \Xi(M_4')} \lambda^e e e^\top = \mathcal{D}\}$, so that $\Lambda(\mathcal{D}) = \{\lambda \in \Lambda_*(\mathcal{D}) \mid \lambda \succeq 0\}$. Recall that elements of $\Xi(M_4')$ are described in Proposition 3.2.6. Since $\#(\Xi(M_4')) = 12 = \dim(\mathcal{S}_4) + 2$, $\Lambda_*(\mathcal{D})$ is a two-dimensional affine space. We compute that $\Lambda_*(\mathcal{D}) = \{\lambda_{\alpha,\beta,\gamma}(\mathcal{D}) \mid \alpha, \beta, \gamma \in \mathbb{R},\ \alpha + \beta + \gamma = 0\}$, from which the result follows. □

In addition to explaining how to compute $\lambda(\mathcal{D})$, the above propositions also allow us to establish, in Theorem 3.3.6 below, the part of Theorem 3.1.3 about Lipschitz regularity of the map λ.

Lemma 3.3.5. *Let $\mathcal{D} \in \mathcal{S}_d^{++}$, and let $M = A^\top M_0 A$ be a minimizing perfect form for $\mathcal{D}$ in (3.4), where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. Then $|A| \leq C\mu(\mathcal{D})$, where $C > 0$ is a constant depending only on the dimension d.*

Proof. It holds that

$$d\lambda_{\max}(\mathcal{D}) \geq \mathrm{Tr}(\mathcal{D}) \geq \mathrm{Tr}(\mathcal{D}A^\top M_0 A) \geq |A|^2 \lambda_{\min}(\mathcal{D})\lambda_{\min}(M_0),$$

where we used the optimality of $M = A^\top M_0 A$ in the second inequality. The result follows. □

Theorem 3.3.6. *Assume that $d \leq 4$, and equip Λ_d with the norm $|\cdot|_\infty\colon \lambda \mapsto \max_{e \in \mathcal{Z}_d} |\lambda^e|$ (or alternatively the natural extension to Λ_d of any norm in $\mathbb{R}^n$). Then the mapping $\mathcal{D} \in \mathcal{S}_d^{++} \mapsto \lambda(\mathcal{D})$ is locally Lipschitz continuous, with dilatation coefficient $K(\mu)$ as defined in Theorem 3.1.3, where $\mu = \mu(\mathcal{D})$.*

Proof. For any perfect form $M \in \mathrm{Perfect}(d)$, we denote by $\mathcal{S}^{++}(M)$ the set of matrices $\mathcal{D} \in \mathcal{S}_d^{++}$ for which M is optimal in (3.4). Note that $\mathcal{S}^{++}(M)$ is characterized by

$$\mathcal{S}^{++}(M) = \{\mathcal{D} \in \mathcal{S}_d^{++} \mid \langle \mathcal{D}, M' - M\rangle \geq 0,\ \forall M' \in \mathcal{N}(M)\}.$$

In particular, $\mathcal{S}^{++}(M)$ is convex.

We deduce easily from Propositions 3.3.3 and 3.3.4 that the mapping λ is Lipschitz continuous on $\bigcup_{M_0 \in \mathrm{Perfect}_0(d)} \mathcal{S}^{++}(M_0)$, with some dilatation coefficient $K_0 > 0$.

Let $\mathcal{D}_1$, $\mathcal{D}_2 \in \mathcal{S}_d^{++}$, $I := \{t\mathcal{D}_1 + (1-t)\mathcal{D}_2 \mid t \in [0,1]\}$, and $\mu := \max\{\mu(\mathcal{D}) \mid \mathcal{D} \in I\} = \max\{\mu(\mathcal{D}_1), \mu(\mathcal{D}_2)\}$. We consider the restriction of the mapping λ to the segment I. If $M = A^\top M_0 A \in \mathrm{Perfect}(d)$, where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$, is such that $I \cap \mathcal{S}^{++}(M)$ is nonempty, then $|A| \leq C\mu$, where $C > 0$ is as in Lemma 3.3.5. Since there are only finitely many $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $|A| \leq C\mu$, it follows that I is the union of finitely many closed segments $I \cap \mathcal{S}^{++}(M)$. By Proposition 3.3.2 and the above, λ is Lipschitz continuous on the segment $I \cap \mathcal{S}^{++}(A^\top M_0 A)$, with dilatation coefficient $K_0|A|^2 \leq K_0 C^2 \mu^2$. Thus λ is $K(\mu)$-Lipschitz on the whole segment I, where $K(\mu) := K_0 C^2 \mu^2$. □

3.4 Upper bound on the radius of the stencil

Our aim in this section is to prove the following result, which implies the part of Theorem 3.1.3 about $R(\mu)$-supportedness (but is not restricted to dimension $d \leq 4$):

Theorem 3.4.1. *For any $\mathcal{D} \in \mathcal{S}_d^{++}$, $\lambda \in \Lambda(\mathcal{D})$, and $e \in \mathcal{Z}_d$, if $\lambda^e > 0$, then*

$$|e| \leq C\mu(\mathcal{D}),$$

where $C > 0$ is a constant depending only on the dimension d.

Theorem 3.4.1 was previously proved in dimension $d \leq 3$, while in higher dimension only the following weaker result was known, see [Mir19, Proposition 1.1]:

Theorem 3.4.2. *For any $\mathcal{D} \in \mathcal{S}_d^{++}$, $\lambda \in \Lambda(\mathcal{D})$, and $e \in \mathcal{Z}_d$, if $\lambda^e > 0$, then*

$$|e| \leq C\mu(\mathcal{D})^{d-1},$$

where $C > 0$ is a constant depending only on the dimension d.

Both the proofs of Theorem 3.4.1 and Theorem 3.4.2 rely on the following lemma.

Lemma 3.4.3. *Let $\mathcal{D} \in \mathcal{S}_d^{++}$, and let $M = A^\top M_0 A$ be a minimizing perfect form for $\mathcal{D}$ in* (3.4)*, where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. For any $\lambda \in \Lambda(\mathcal{D})$ and $e \in \mathcal{Z}_d$, if $\lambda^e > 0$, then $|e| \leq C|A^{-1}|$, where $C > 0$ is a constant depending only on the dimension d.*

Proof. By Proposition 3.2.3, one has $e \in \Xi(M)$, hence $Ae \in \Xi(M_0)$. The result follows, with $C = \max_{M_0 \in \mathrm{Perfect}_0(d)} \max_{e \in \Xi(M_0)} |e|$. □

Theorem 3.4.2 follows from Lemma 3.4.3, Lemma 3.3.5, and from the fact that for any $A \in \mathrm{GL}_d(\mathbb{Z})$, one has $|A^{-1}| \leq |A|^{d-1}$, since $1 = |\det(A)| \leq |A|^{d-1}|A^{-1}|^{-1}$. However, Lemma 3.3.5 is not sufficient to conclude the proof of Theorem 3.4.1.

The bulk of this section is devoted to proving the following lemma:

Lemma 3.4.4. *Let $M_0 \in \mathrm{Perfect}_0(d)$, and let $\mathcal{S}^{++}(M_0)$ denote the set of matrices $\mathcal{D} \in \mathcal{S}_d^{++}$ for which M_0 is optimal in* (3.4)*. There exists a finite set $\mathrm{Perfect}_1(d; M_0) \subset \mathrm{Perfect}(d)$ such that for any $\mathcal{D} \in \mathcal{S}_d^{++}$ satisfying $\mathcal{D}^{-1} \in \mathcal{S}^{++}(M_0)$, some $M_1 \in \mathrm{Perfect}_1(d; M_0)$ is optimal in* (3.4)*.*

For convenience, we denote

$$\mathrm{Perfect}_1(d) := \bigcup_{M_0 \in \mathrm{Perfect}_0(d)} \mathrm{Perfect}_1(d; M_0).$$

Assuming Lemma 3.4.4, we may conclude the proof of the main result:

Proof of Theorem 3.4.1. Denote by $M = A^\top M_0 A$ a minimizing perfect form for $\mathcal{D} \in \mathcal{S}_d^{++}$ in (3.4), where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. Then M_0 is a minimizing perfect form for $A\mathcal{D}A^\top$ (see Proposition 3.3.2). By Lemma 3.4.4, some $M_1 \in \mathrm{Perfect}_1(d)$ is a minimizing perfect form for $A^{-\top}\mathcal{D}^{-1}A^{-1}$, and finally $A^{-1}M_1A^{-\top}$ is a minimizing perfect form for $\mathcal{D}^{-1}$.

We use a variant of Lemma 3.3.5, where the finite set $\mathrm{Perfect}_0(d)$ is replaced by $\mathrm{Perfect}_1(d)$: it holds that

$$d\lambda_{\max}(\mathcal{D}^{-1}) \geq \mathrm{Tr}(\mathcal{D}^{-1}) \geq \mathrm{Tr}(\mathcal{D}^{-1}A^{-1}M_1A^{-\top}) \geq |A^{-1}|^2\lambda_{\min}(\mathcal{D}^{-1})\lambda_{\min}(M_1),$$

therefore $|A^{-1}| \leq C\mu(\mathcal{D}^{-1}) = C\mu(\mathcal{D})$, where $C > 0$ depends only on d. We conclude using Lemma 3.4.3. □

One also has the following estimate, which is sharper than Theorem 3.4.1 since $|e|_{\mathcal{D}^{-1}} \geq |e|\lambda_{\min}(\mathcal{D}^{-1})^{1/2} = |e|\lambda_{\max}(\mathcal{D})^{-1/2}$:

Corollary 3.4.5. *For any $\mathcal{D} \in \mathcal{S}_d^{++}$, $\lambda \in \Lambda(\mathcal{D})$, and $e \in \mathcal{Z}_d$, if $\lambda^e > 0$, then*

$$|e|_{\mathcal{D}^{-1}} \leq C\lambda_{\min}(\mathcal{D})^{-1/2},$$

where $C > 0$ is a constant depending only on the dimension d.

Proof. From the previous argument, we obtain

$$|\mathcal{D}^{-1/2}A^{-1}|^2 \leq \mathrm{Tr}(A^{-\top}\mathcal{D}^{-1}A^{-1}) \leq C_0/\lambda_{\min}(\mathcal{D}),$$

where $C_0 > 0$ depends only on d. Therefore

$$|Ae| = |A\mathcal{D}^{1/2}\mathcal{D}^{-1/2}e| \geq |\mathcal{D}^{-1/2}A^{-1}|^{-1}|e|_{\mathcal{D}^{-1}} \geq (\lambda_{\min}(\mathcal{D})/C_0)^{1/2}|e|_{\mathcal{D}^{-1}}.$$

Since $Ae \in \Xi(M_0)$, one has $|Ae| \leq C_1$, where $C_1 > 0$ depends only on d. This concludes the proof. □

Let us now turn to the proof of Lemma 3.4.4. For any matrix $A \in \mathbb{R}^{d\times d}$, we denote by $\mathrm{adj}(A)$ its adjugate matrix (if A is invertible, then $A^{-1} = \det(A)^{-1}\,\mathrm{adj}(A)$ by Cramer's rule). For any $E \subset \mathcal{Z}_d$, we let $\mathcal{D}_E := \mathrm{adj}(\sum_{e\in E} ee^\top)$.

Lemma 3.4.6. *Let $M_0 \in \mathrm{Perfect}_0(d)$. Then any $\mathcal{D} \in \mathcal{S}_d^{++}$ for which M_0 is optimal in* (3.4) *satisfies*

$$\mathcal{D}^{-1} \in \mathrm{Cone}\,\{\mathcal{D}_E \mid E \subset \Xi(M_0)\},$$

where Cone *denotes the convex conical hull.*

Proof. For any $A \in \mathbb{R}^{d\times d}$, u, $v \in \mathbb{R}^d$, and $t \in [0,1]$, it holds [Dac08, Proposition 5.65] that

$$\mathrm{adj}(A + tuv^\top) = t\,\mathrm{adj}(A + uv^\top) + (1-t)\,\mathrm{adj}(A).$$

In our setting, by Proposition 3.2.3, there are weights $(\lambda^e)_{e\in\Xi(M_0)}$ such that $\mathcal{D} = \sum_{e\in\Xi(M_0)} ee^\top$, and we may assume up to rescaling $\mathcal{D}$ that $\lambda^e \in [0,1]$, for any $e \in \Xi(M_0)$. Applying the above formula recursively to $\mathrm{adj}(\mathcal{D}) = \det(\mathcal{D})\mathcal{D}^{-1}$ yields

$$\begin{aligned}\mathrm{adj}(\mathcal{D}) &= \sum_{E\subset\Xi(M_0)} \left(\prod_{e\in E}\lambda^e\right)\left(\prod_{e\in\Xi(M_0)\setminus E}(1-\lambda^e)\right)\mathrm{adj}\left(\sum_{e\in E} ee^\top\right)\\ &= \sum_{E\subset\Xi(M_0)} \left(\prod_{e\in E}\lambda^e\right)\left(\prod_{e\in\Xi(M_0)\setminus E}(1-\lambda^e)\right)\mathcal{D}_E,\end{aligned}$$

which concludes the proof. □

For any $M_0 \in \mathrm{Perfect}_0(d)$, let us denote by $\mathcal{E}(M_0)$ the set of parts of $\Xi(M_0)$, and let $\mathcal{L}(M_0) \subset \mathbb{R}_+^{\mathcal{E}(M_0)}$ be defined by

$$\mathcal{L}(M_0) := \{(\mathrm{Tr}(\mathcal{D}_E M))_{E \in \mathcal{E}(M_0)} \mid M \in \mathrm{Perfect}(d)\}.$$

Lemma 3.4.7. *There exists $n_0(d) \in \mathbb{N}^*$ such that for any $M_0 \in \mathrm{Perfect}_0(d)$, $\mathcal{L}(M_0) \subset \frac{1}{n_0(d)}\mathbb{N}^{\mathcal{E}(M_0)}$.*

Proof. The elements of $\mathrm{Perfect}_0(d)$ have rational coefficients, since they are the vertices of a polytope defined by rational inequalities. Also, $\mathcal{D}_E$ and elements of $\mathrm{GL}_d(\mathbb{Z})$ have integer coefficients. Therefore, it suffices to choose $n_0(d)$ such that $n_0(d)M_0$ has integer coefficients for any $M_0 \in \mathrm{Perfect}_0(d)$. □

We equip $\mathbb{R}_+^{\mathcal{E}(M_0)}$ with the componentwise partial ordering.

Corollary 3.4.8. *For any $M_0 \in \mathrm{Perfect}_0(d)$, the set of minimal elements of $\mathcal{L}(M_0)$, denoted $\mathcal{L}_1(M_0)$, is finite.*

Proof. It suffices to prove that the set of minimal elements of any $A \subset \mathbb{N}^N$ is finite, where N is arbitrary. Hence, it suffices to prove that there is no sequence $(\alpha_k)_{k\geq 0}$ of pairwise non-comparable elements of $\mathbb{N}^N$. For contradiction, consider such a sequence. Then for any $k \geq 0$, there exists $1 \leq i \leq N$ such that $a_k[i] < a_0[i]$. Thus, there exists $1 \leq i \leq N$ and a subsequence such that $a_{\sigma(k)}[i]$ is independent of k. But this produces an infinite sequence of pairwise non-comparable elements in $\mathbb{N}^{N-1}$, which by induction yields a contradiction (the case $N = 1$ being obvious), and the result is proved. □

Proof of Lemma 3.4.4. We let $\mathrm{Perfect}_1(d; M_0) \subset \mathrm{Perfect}(d)$ be a set of antecedents of $\mathcal{L}_1(M_0) \subset \mathcal{L}(M_0)$, with $\#(\mathrm{Perfect}_1(d; M_0)) = \#(\mathcal{L}_1(M_0)) < +\infty$, by Corollary 3.4.8. Let $\mathcal{D} \in \mathcal{S}_d^{++}$ be such that $M_0 \in \mathrm{Perfect}_0(d)$ solves (3.4). By Lemma 3.4.6, there exist weights $\mu \colon \mathcal{E} \to \mathbb{R}_+$ such that

$$\mathcal{D}^{-1} = \sum_{E \subset \mathcal{E}(M_0)} \mu(E)\mathcal{D}_E,$$

and therefore, for any $M \in \mathrm{Perfect}(d)$,

$$\mathrm{Tr}(\mathcal{D}^{-1}M) = \sum_{E \in \mathcal{E}(M_0)} \mu(E)\,\mathrm{Tr}(\mathcal{D}_E M).$$

By construction, the above quantity is minimized for some $M \in \mathrm{Perfect}_1(d; M_0)$. □

3.4.1 Conjectured constructive variant of Lemma 3.4.4

The above proof is not constructive; however, we illustrate it with the following numerical experiment. For any $M_0 \in \mathrm{Perfect}_0(d)$, we denote by $\mathrm{Perfect}_1^*(d; M_0)$ the set of perfect forms $M \in \mathrm{Perfect}(d)$ that are minimizing in (3.4) for $\mathcal{D} = (\sum_{e \in \Xi(M_0)} ee^\top)^{-1}$, and by $\mathrm{Perfect}'(d; M_0)$ the set of perfect forms of $\mathcal{M}_d$ whose distance to $\mathrm{Perfect}_1^*(d; M_0)$ on the graph of vertices of $\mathcal{M}_d$ is less than or equal to four. We define $\mathcal{L}'(M_0) \subset \mathbb{R}_+^{\mathcal{E}(M_0)}$ by

$$\mathcal{L}'(M_0) := \{(\mathrm{Tr}(\mathcal{D}_E M))_{E \in \mathcal{E}(M_0)} \mid M \in \mathrm{Perfect}'(d; M_0)\},$$

and we denote by $\mathcal{L}'_1(M_0)$ the set of minimizers of $\mathcal{L}'(M_0)$ and by $\mathrm{Perfect}'_1(d; M_0) \subset \mathrm{Perfect}'(d; M_0)$ the set of all antecedents of $\mathcal{L}'_1(M_0)$. Note that by construction, if $\mathrm{Perfect}_1(d; M_0) \subset \mathrm{Perfect}'(d; M_0)$, then $\mathrm{Perfect}_1(d; M_0) \subset \mathrm{Perfect}'_1(d; M_0)$. We conjecture that those inclusions are satisfied, yielding the following variant of Lemma 3.4.4:

Conjecture 3.4.9. *Let $d \in \{2,3,4\}$, let $M_0 \in \mathrm{Perfect}_0(d)$, and let $\mathcal{S}^{++}(M_0)$ denote the set of matrices $\mathcal{D} \in \mathcal{S}_d^{++}$ for which M_0 is optimal in* (3.4)*. Then for any $\mathcal{D} \in \mathcal{S}_d^{++}$ satisfying $\mathcal{D}^{-1} \in \mathcal{S}^{++}(M_0)$, some $M_1 \in \mathrm{Perfect}_1'(d; M_0)$ is optimal in* (3.4)*.*

We observe that

$$\#(\mathrm{Perfect}_1'(2; M_2^*)) = 1, \qquad \#(\mathrm{Perfect}_1'(3; M_3^*)) = 3,$$
$$\#(\mathrm{Perfect}_1'(4; M_4^*)) = 22, \qquad \#(\mathrm{Perfect}_1'(4; M_4')) = 545,$$

and that $\mathrm{Perfect}_1'(d; M_d^*)$ coincides with $\mathrm{Perfect}_1^*(d; M_d^*)$ for any $2 \le d \le 4$, while elements of $\mathrm{Perfect}_1'(4; M_4')$ are at distance at most two to $\mathrm{Perfect}_1^*(4; M_4')$ on the graph of vertices of $\mathcal{M}_4$. We also observe that

$$\mathrm{Perfect}_1'(2; M_2^*) = \left\{ \frac{1}{2} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \right\},$$

that elements of $\mathrm{Perfect}_1'(3; M_3^*)$ are of the form

$$\frac{1}{2}(3I_3 - \mathbb{1}\mathbb{1}^\top + e_i e_j^\top + e_j e_i^\top) = \frac{1}{2}\left(\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} + e_i e_j^\top + e_j e_i^\top \right), \quad 1 \le i < j \le 3,$$

where $\mathbb{1} := (1, \dots, 1)$ and $(e_1, \dots, e_d)$ denotes the canonical basis of $\mathbb{R}^d$, and that elements of $\mathrm{Perfect}_1'(4; M_4^*)$ are of either of the following forms:

$$\frac{1}{2}(3I_4 - \mathbb{1}\mathbb{1}^\top + e_i e_j^\top + e_j e_i^\top + e_j e_k^\top + e_k e_j^\top + e_k e_l^\top + e_l e_k^\top), \quad \{i,j,k,l\} = \{1,2,3,4\},$$
$$\frac{1}{2}(3I_4 - \mathbb{1}\mathbb{1}^\top + e_i e_j^\top + e_j e_i^\top + e_i e_k^\top + e_k e_i^\top + e_j e_k^\top + e_k e_j^\top), \quad \{i,j,k,l\} = \{1,2,3,4\},$$
$$\frac{1}{2}(3I_4 - \mathbb{1}\mathbb{1}^\top + e_i e_j^\top + e_j e_i^\top + 2e_k e_l^\top + 2e_l e_k^\top), \quad \{i,j,k,l\} = \{1,2,3,4\}.$$

If Conjecture 3.4.9 is true, then the proofs of Theorem 3.4.1 and Corollary 3.4.5 yield the value

$$\max_{\substack{M_0 \in \mathrm{Perfect}_0(d) \\ M_1 \in \mathrm{Perfect}_1'(d; M_0)}} \frac{\sqrt{d}\, \max_{e \in \Xi(M_0)} |e|}{\sqrt{\lambda_{\min}(M_1)}}$$

for the constant C. By computing this value for the dimension $d = 2$ (respectively $d = 3$, $d = 4$), we obtain the conjecture that $C = 2\sqrt{2} \approx 2.828$ (respectively $C = \sqrt{12 + 6\sqrt{2}} \approx 4.526$, $C = 2\sqrt{34 + 2\sqrt{281}} \approx 16.435$) in this dimension. These estimates are not sharp; in dimension $d = 2$, the sharper estimate $C = 2$ was proved in [Mir18, Theorem 4.11].

For completeness, we mention that the sets $\mathrm{Perfect}'(2; M_2^*)$, $\mathrm{Perfect}'(3; M_3^*)$, $\mathrm{Perfect}'(4; M_4^*)$, $\mathrm{Perfect}'(4; M_4')$, and $\mathrm{Perfect}_1^*(4; M_4')$ (which are intermediate results in the process of computing $\mathrm{Perfect}_1'(d; M_0)$ for respectively $d = 2$, 3, 4, 4, 4 and $M_0 = M_2^*$, M_3^*, M_4^*, M_4', M_4') respectively have cardinalities 46, 631, 464329, 92393, and 1.

3.5 Guarantees against checkerboard artifacts

We establish here the part of Theorem 3.1.3 about the ε-spanning property:

Theorem 3.5.1. *Assume that $d \le 4$. For any $\mathcal{D} \in \mathcal{S}_d^{++}$,*

$$\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e(\mathcal{D}) \ge \varepsilon |\mathcal{D}^{-1}|^{-1}\} = \mathbb{Z}^d,$$

where $\varepsilon > 0$ is a constant depending only on d.

Theorem 3.5.1 is a new result in dimension $d = 4$; in dimensions $d \in \{2, 3\}$, it was established and used for the first time in the context in numerical analysis in Chapter 4, section 4.4.3.

Lemma 3.5.2. *For any $\mathcal{D} \in \mathcal{S}_d^{++}$ and any $\lambda \in \Lambda(\mathcal{D})$,*

$$\mathrm{Span}_{\mathbb{R}}\{e \in \mathcal{Z}_d \mid \lambda^e \geq \varepsilon |\mathcal{D}^{-1}|^{-1}\} = \mathbb{R}^d,$$

where $\varepsilon > 0$ is a constant depending only on d.

Proof. Let $n_d := \max_{M_0 \in \mathrm{Perfect}_0(d)} \#(\Xi(M_0))$, and let

$$c_d := \min \left\{ \max_{i_1 < \cdots < i_d} \lambda_{\min} \left(\sum_{i=1}^{d} e_{i_j} e_{i_j}^\top \right) \;\middle|\; (e_1, \ldots, e_{n_d}) \in (\mathbb{R}^d)^{n_d}, \sum_{i=1}^{n_d} e_i = I_d \right\}.$$

Then c_d is positive, as the minimum of a positive continuous function over a compact set, and it depends only on d. By a simple change of variables, for any $(e_1, \ldots, e_{n_d}) \in (\mathbb{R}^d)^{n_d}$ such that $\sum_{i=1}^{n_d} e_i e_i^\top = \mathcal{D}$, there exists $i_1 < \cdots < i_d$ such that $\sum_{j=1}^{d} e_{i_j} e_{i_j}^\top \succeq c_d \mathcal{D}$.

Let $M \in \mathrm{Perfect}(\mathcal{D})$ be minimizing in (3.4). By Proposition 3.2.3, one has $\mathcal{D} = \sum_{e \in \Xi(M)} \lambda^e e e^\top$, where $\#(\Xi(M)) \leq n_d$. Thus there exists $\Xi \subset \Xi(M)$, $\#(\Xi) = d$, such that $\sum_{e \in \Xi} \lambda^e e e^\top \succeq c_d \mathcal{D}$ (and therefore $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$).

Let $e \in \Xi$, and let $v \in \mathbb{R}^d \setminus \{0\}$ be orthogonal to $\mathrm{Span}_{\mathbb{R}}(\Xi \setminus \{e\})$. Using Corollary 3.4.5 for the last inequality, one has

$$c_d |v|_{\mathcal{D}}^2 \leq \sum_{e' \in \Xi} \lambda^{e'} \langle e', v \rangle^2 = \lambda^e \langle e, v \rangle^2 \leq \lambda^e |e|_{\mathcal{D}^{-1}}^2 |v|_{\mathcal{D}}^2 \leq \lambda^e C |\mathcal{D}^{-1}| |v|_{\mathcal{D}}^2,$$

where $C > 0$ depends only on d. Therefore $\lambda^e \geq (c_d / C) |\mathcal{D}^{-1}|^{-1}$, which concludes the proof. □

Lemma 3.5.3. *Let $d \leq 4$ and $M = A^\top M_0 A \in \mathrm{Perfect}(d)$, where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. If $M_0 = M_d^*$, then any d-element subset $\Xi \subset \Xi(M)$ satisfying $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$ also satisfies $\mathrm{Span}_{\mathbb{Z}} \Xi = \mathbb{Z}^d$. This remains true if $M_0 = M_4'$, except for subsets $\{A^{-1} e \mid e \in \Xi_i\} \subset \Xi(M)$, $1 \leq i \leq 3$, where*

$$\begin{aligned}
\Xi_1 &:= \{\pm e_1, \pm e_2, \pm(e_2 - e_3), \pm(e_1 - e_2 - e_3 + e_4)\}, \\
\Xi_2 &:= \{\pm e_2, \pm(e_1 - e_3), \pm(e_4 - e_3), \pm(e_1 - e_2 + e_4)\}, \\
\Xi_3 &:= \{\pm e_3, \pm(e_1 - e_2), \pm(e_4 - e_2), \pm(e_1 - e_3 + e_4)\}.
\end{aligned}$$

Proof. By Proposition 3.2.4, we may assume without loss of generality that $A = I_d$. Then the proof is by exhaustive computer enumeration (denoting by C_n^k the binomial coefficients, one has to enumerate the $C^4_{\Xi(M_4^*)} = C^4_{10} = 210$ subsets of $\Xi(M_4^*)$, and the $C^4_{\Xi(M_4')} = C^4_{12} = 495$ subsets of $\Xi(M_4')$). □

Corollary 3.5.4. *There exists $\mathcal{D} \in \mathcal{S}_4^{++}$ and $\lambda \in \Lambda(\mathcal{D})$ such that $\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e > 0\} \neq \mathbb{Z}^d$.*

Proof. Choose $\mathcal{D} = \sum_{e \in \Xi_1} e e^\top$ and $\lambda^e := 1$ if $e \in \Xi_1$, $\lambda^e := 0$ otherwise. □

The above corollary shows that it is important to choose $\lambda(\mathcal{D})$ as the barycenter of $\Lambda(\mathcal{D})$, and not just any point of $\Lambda(\mathcal{D})$, in order for Theorem 3.5.1 to apply.

Proof of Theorem 3.5.1. Let $M = A^\top M_0 A \in \mathrm{Perfect}(d)$ be minimizing in (3.4), where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. For now, let ε be as in Lemma 3.5.2. Then there exists $\Xi \subset \Xi(M)$, $\#(\Xi) = d$, such that $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$ and $\lambda^e(\mathcal{D}) \geq \varepsilon |\mathcal{D}^{-1}|^{-1}$, for any $e \in \Xi$.

We assume from now on that $M_0 = M_4'$ and $\Xi = \{A^{-1}e \mid e \in \Xi_i\}$, for some $1 \leq i \leq 3$, since otherwise Lemma 3.5.3 concludes the proof. Let $\kappa_1 := \min\{\lambda^e(\mathcal{D}) \mid e \in \Xi\}$ and $\kappa_2 := \min\{\lambda^e(\mathcal{D}) \mid e \in \Xi(M)\}$. We know that $\kappa_1 \geq \varepsilon |\mathcal{D}^{-1}|^{-1}$. Let us show that $\kappa_2 \geq (\varepsilon/4)|\mathcal{D}^{-1}|^{-1}$.

Note that $\Xi(M_4') = \Xi_1 \cup \Xi_2 \cup \Xi_3$ and that $\sum_{e \in \Xi_i} e e^\top$ is independent of $i \in \{1,2,3\}$ (thus so is $\sum_{e \in \Xi_i} (A^{-1}e)(A^{-1}e)^\top$). We deduce that $\tilde{\lambda} \in \Lambda(\mathcal{D})$, where

$$\tilde{\lambda}^e := \begin{cases} \lambda^e(\mathcal{D}) - \kappa_1 & \text{if } e \in \Xi, \\ \lambda^e(\mathcal{D}) + \kappa_1/2 & \text{if } e \in \Xi(M) \setminus \Xi, \\ 0 & \text{else.} \end{cases}$$

Since $\Lambda(\mathcal{D})$ is an equilateral triangle and $\lambda(\mathcal{D})$ is its barycenter, the point $\hat{\lambda} := (3/2)\lambda(\mathcal{D}) - (1/2)\tilde{\lambda}$ belongs to $\Lambda(\mathcal{D})$. By construction, there is $e_* \in \Xi(M) \setminus \Xi$ such that $\kappa_2 = \lambda^{e_*}(\mathcal{D})$. One has

$$0 \leq \hat{\lambda}^{e_*} = \kappa_2 - \kappa_1/4.$$

Thus, for any $e \in \Xi(M)$, one has $\lambda^e(\mathcal{D}) \geq \kappa_2 \geq \kappa_1/4 \geq (\varepsilon/4)|\mathcal{D}^{-1}|^{-1}$. This concludes the proof, since $\mathrm{Span}_{\mathbb{Z}} \Xi(M) = \mathrm{Span}_{\mathbb{Z}} \Xi(M_4') = \mathbb{Z}^d$. □

The spanning property prevents checkerboard artifacts by ensuring that the connectivity of finite difference discretizations, as illustrated by the following proposition:

Proposition 3.5.5. *Assume that $d \leq 4$. Let $\mathcal{D} \colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ be K-Lipschitz and have bounded condition number and uniformly positive smallest eigenvalue. Define $\lambda^e(x) := \lambda^e(\mathcal{D}(x))$, for any $x \in \mathbb{R}^d$ and $e \in \mathcal{Z}_d$, where λ is from Definition 3.1.2. Then for $h > 0$ small enough and for any x, $y \in h\mathbb{Z}^d$, there exist $n \in \mathbb{N}$, $n \leq C|x-y|/h$, and a family $(x_i)_{0 \leq i \leq n}$ of points of $h\mathbb{Z}^d$ such that $x_0 = x$, $x_n = y$, and for any $0 \leq i < n$, $\lambda^{(x_{i+1}-x_i)/h}(x_i) \geq \varepsilon$, where C, $\varepsilon > 0$ depend only on d, K, $\|\mu(\mathcal{D})\|_\infty$, and $\|\mathcal{D}^{-1}\|_\infty$.*

Proof. By Theorem 3.5.1, there exists $\Xi \subset \mathcal{Z}_d$, $\#(\Xi) = d$, such that $\mathrm{Span}_{\mathbb{Z}} \Xi = \mathbb{Z}^d$ and, for any $e \in \Xi$, $\lambda^e(x) \geq 2\varepsilon$, where $\varepsilon > 0$ depends only on d and $\|\mathcal{D}^{-1}\|_\infty$. Denote by $E \in \mathbb{R}^{d \times d}$ a matrix whose columns are elements of Ξ. Then $E \in \mathrm{GL}_d(\mathbb{Z})$.

Let e_i, $1 \leq i \leq d$ be a vector of the canonical basis of $\mathbb{R}^d$. There exist $(\nu_e)_{e \in \Xi} \in \mathbb{Z}^\Xi$ such that $\sum_{e \in \Xi} \nu_e e = e_i$ and ν_e are elements of a row of the matrix E^{-1}. Hence $\sum_{e \in \Xi} |\nu_e| \leq |E^{-1}|_\infty$.

Let $M = A^\top M_0 A \in \mathrm{Perfect}(d)$ be minimizing in (3.4) for $\mathcal{D} = \mathcal{D}(x)$, where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. By Proposition 3.2.3, one has $\Xi \subset \Xi(M)$. By Proposition 3.2.4, one has $E = A^{-1}E_0$, where $E_0 \in \mathrm{GL}_d(\mathbb{Z})$ is a matrix whose columns are elements of $\Xi(M_0)$. By Lemma 3.3.5, $|E^{-1}|_\infty = |E_0^{-1}A|_\infty \leq C_0$, where $C_0 > 0$ depends only on d and $\|\mu(\mathcal{D})\|_\infty$. Hence $\sum_{e \in \Xi} |\nu_e| \leq C_0$.

Let $n := \sum_{e \in \Xi} |\nu_e|$. Then there exists a family $(x_i)_{0 \leq i \leq n}$ of points of $h\mathbb{Z}^d$ such that $x_0 = x$, $x_n = x + he_i$, and for any $0 \leq i < n$, $\pm(x_{i+1} - x_i)/h \in \Xi$. By Theorem 3.4.1, there exists $R > 0$, depending only on d and $\|\mu(\mathcal{D})\|_\infty$, such that $|x_i - x| \leq C_0 R h$, for any $0 \leq i \leq n$. By Theorem 3.3.6, there exists $K' > 0$, depending only on d, K, and $\|\mu(\mathcal{D})\|_\infty$, such that $\lambda^e(x) \geq 2\varepsilon - K' C_0 R h$, for any $e \in \Xi$ and $0 \leq i \leq n$. We choose $h \leq \varepsilon/(K' C_0 R)$, so that $2\varepsilon - K' C_0 R h \geq \varepsilon$.

We conclude the proof by concatenating families $(x_i)_{0 \leq i \leq n}$ built as above and starting at points x, $x + he_{i_1}$, $x + he_{i_1} + he_{i_2}$, …, $x + he_{i_1} + \cdots + he_{i_{k-1}}$, where e_{i_1}, …, e_{i_k} are vectors belonging to the canonical basis of $\mathbb{R}^d$ up to a change of sign, satisfying $x + he_{i_1} + \cdots + he_{i_k} = y$ and $k = |y - x|_1/h$. □

The following is an example of a finite difference scheme featuring checkerboard artifacts:

$$\frac{1}{2}\sum_{i=1}^{2}\frac{u(x+he_i)+u(x-he_i)-2u(x)}{h^2}=0 \quad \text{in } (h\mathbb{Z}^2)/\mathbb{Z}^2,$$

where $e_1 := (1,1)$, $e_2 := (1,-1)$, and $h > 0$, $1/h \in 2\mathbb{N}$. This scheme is a discretization of the equation $\Delta u(x) = 0$ in $\mathbb{R}^2/\mathbb{Z}^2$. Its set of solutions is the set of functions that are constant on each one of the two distinct sets

$$(h\{(i,j)\in\mathbb{Z}^2 \mid i+j\in 2\mathbb{Z}\})/\mathbb{Z}^2, \qquad (h\{(i,j)\in\mathbb{Z}^2 \mid i+j\in 2\mathbb{Z}+1\})/\mathbb{Z}^2$$

of points of $(h\mathbb{Z}^2)/\mathbb{Z}^2$ whose sums of coefficients are respectively even and odd after scaling by h^{-1}, but that are not necessarily constant on the whole lattice $(h\mathbb{Z}^2)/\mathbb{Z}^2$.

We show below that some finite difference schemes of the form

$$\sum_{e\in\mathcal{Z}_d}\lambda^e(x)\frac{u(x+he)+u(x-he)-2u(x)}{h^2}=0 \quad \text{in } (h\mathbb{Z}^d)/\mathbb{Z}^d,$$

where the left-hand side is built using the approximation (3.1), are guaranteed not to feature such checkerboard artifacts, in the sense that their solutions are always constant functions on the whole lattice $(h\mathbb{Z}^d)/\mathbb{Z}^d$. The following corollary is a quantitative variant of this statement. Attempting to prove an improved estimate without the factor $\exp(1/h)$ in the right-hand side is an opportunity for future work.

Corollary 3.5.6. *Assume that $d \le 4$. Let $\mathcal{D}\colon (\mathbb{R}^d \setminus \mathbb{Z}^d) \to \mathcal{S}_d^{++}$ be K-Lipschitz continuous. Define $\lambda^e(x) := \lambda^e(\mathcal{D}(x))$, for any $x \in \mathbb{R}^d$ and $e \in \mathcal{Z}_d$, where λ is from Definition 3.1.2. Let $h > 0$ be small enough and such that $1/h \in \mathbb{N}$. Then for any $u\colon (h\mathbb{Z}^d)/\mathbb{Z}^d \to \mathbb{R}$,*

$$\max_{x,\,y\in(h\mathbb{Z}^d)/\mathbb{Z}^d}|u(x)-u(y)| \le \exp(C/h)\max_{x\in(h\mathbb{Z}^d/\mathbb{Z}^d)}\left|\sum_{e\in\mathcal{Z}_d}\lambda^e(x)\frac{u(x+he)+u(x-he)-2u(x)}{h^2}\right|,$$

where $C > 0$ is a constant depending only on d, K, $\|\mathcal{D}\|_\infty$, and $\|\mathcal{D}^{-1}\|_\infty$.

Proof. We may assume without loss of generality that $\min_{x\in(h\mathbb{Z}^d)/\mathbb{Z}^d} u(x) = 0$, up to adding some constant to the function u.

Let

$$C_0 := \max_{x\in(h\mathbb{Z}^d/\mathbb{Z}^d)}\left|\sum_{e\in\mathcal{Z}_d}\lambda^e(x)\frac{u(x+he)+u(x-he)-2u(x)}{h^2}\right|.$$

Then for any $x \in (h\mathbb{Z}^d)/\mathbb{Z}^d$ and $e \in \mathbb{Z}^d \setminus \{0\}$,

$$\begin{aligned}\frac{\lambda^e(x)}{h^2}u(x+he) &\le C_0 - \frac{\lambda^e(x)}{h^2}u(x-he) - \sum_{e'\in\mathcal{Z}_d\setminus\{\pm e\}}\frac{\lambda^{e'}(x)}{h^2}(u(x+he')+u(x-he')) \\ &\quad + 2\sum_{e'\in\mathcal{Z}_d}\frac{\lambda^{e'}(x)}{h^2}u(x) \\ &\le C_0 + 2\sum_{e'\in\mathcal{Z}_d}\frac{\lambda^{e'}(x)}{h^2}u(x) \le C_0 + \frac{C_1}{h^2}u(x),\end{aligned}$$

where $C_1 > 0$ depends only on $\|\mathcal{D}\|_\infty$, where we used that $\sum_{e' \in \mathcal{Z}_d} \lambda^{e'}(x) \leq \sum_{e' \in \mathcal{Z}_d} \lambda^{e'}(x)|e'|^2 = \mathrm{Tr}(\mathcal{D}(x))$. If $\lambda^e(x) > 0$, this yields

$$u(x + he) \leq \frac{h^2 C_0}{\lambda^e(x)} + \frac{C_1}{\lambda^e(x)} u(x). \tag{3.12}$$

Let x, $y \in (h\mathbb{Z}^d)/\mathbb{Z}^d$ be such that

$$u(y) - u(x) = \max_{x',\, y' \in (h\mathbb{Z}^d)/\mathbb{Z}^d} |u(x') - u(y')|.$$

Then $u(x) = 0$, and by Proposition 3.5.5, there exist $n \in \mathbb{N}$, $n \leq C_2/h$, and a family $(x_i)_{0 \leq i \leq n}$ of points of $(h\mathbb{Z}^d) \setminus \mathbb{Z}^d$ such that $x_0 = x$, $x_n = y$, and $\lambda^{(x_{i+1} - x_i)/h}(x_{i_*}) \geq \varepsilon$, for any $0 \leq i < n$, where C_2, $\varepsilon > 0$ depend only on d, K, $\|\mu(\mathcal{D})\|_\infty$, and $\|\mathcal{D}^{-1}\|_\infty$. By (3.12), for any $0 \leq i < n$,

$$u(x_{i+1}) \leq \frac{h^2 C_0}{\varepsilon} + \frac{C_1}{\varepsilon} u(x_i).$$

We conclude using the discrete Grönwall inequality. □

3.A Computing the adjacency relations between perfect forms

Both in the proof of Proposition 3.2.9 and in order to apply Algorithm 3, we need, for any perfect form $M \in \mathrm{Perfect}_0(d)$, to compute an expression of its neighbors $M' \in \mathcal{N}(M)$ in the form $M' = A^\top M_0 A$, where $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$. This procedure, known as Voronoi's algorithm [Vor08], has been extensively studied, and needs to be highly optimized in order to remain practical in large dimensions (such as $d = 8$, see for instance [Sch09b]), while in smaller dimensions it has historically been applied without computer assistance [Mar03], using symmetries of the problem that we do not discuss in this chapter. For information purposes, we describe in this appendix a simple computer assisted strategy to implement Voronoi's algorithm, that remains practical in the dimensions considered in this chapter, that is, for $d \leq 4$.

Let $M \in \mathrm{Perfect}_0(d)$. The procedure that we use in order to characterize its neighbors may be summarized as follows:

1. Compute $\Xi(M)$ (remember that by Lemma 3.2.7, it suffices to check whether finitely many vectors $e \in \mathcal{Z}_d$ satisfy $\langle e, Me \rangle = 1$).

2. For any set $E \subset \Xi(M)$ of size $d(d+1)/2 - 1$ such that symmetric matrices $(ee^\top)_{e \in E}$ are linearly independent, find $P \in \mathcal{S}_d \setminus \{0\}$ such that $\langle e, Pe \rangle = 0$, for any $e \in E$. Then P and $-P$ are *potentially* directions of edges of $\mathcal{M}_d$ stemming from the perfect form M.

3. Up to changing its sign, check whether P is really the direction of such an edge, that is, check whether $\langle e, Pe \rangle \geq 0$, for any $e \in \Xi(M)$. If this property is satisfied, then P is the direction of an edge: continue with step 4. Otherwise, go back to step 2 and continue with the next subset $E \subset \Xi(M)$.

4. Check whether there exists $e' \in \mathcal{Z}_d$ such that $e'(e')^\top$ is linearly independent of $(ee^\top)_{e \in E}$ and such that the matrix $M' \in \mathcal{S}_d$ uniquely determined by $\langle e, M'e \rangle = 1$ for any $e \in E \cap \{e'\}$ belongs to $\mathcal{M}_d$ and is different from M. If such an e' exists, then the associated M' belongs to $\mathcal{N}(M)$ and is the neighbor vertex to M in the direction P: continue with step 5. Otherwise, P is the direction of an unbounded edge. We observe that this never happens in practice, proving that no unbounded edge stem from the vertex M of $\mathcal{M}_d$.

5. Find $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M_0 \in \mathrm{Perfect}_0(d)$ such that $M' = A^\top M_0 A$.

We need to give some details about the implementation of steps 4 and 5.

At step 4, in order to find the suitable vector $e' \in \mathcal{Z}_d$, we simply iterate over all $e' \in \mathcal{Z}_d$, by order of increasing norm, until e' satisfies the required properties. We need to explain how we check whether the associated matrix M' belongs to $\mathcal{M}_d$. We proceed in two steps. First we check whether $M' \in \mathcal{S}_d^{++}$, by attempting to compute its Cholesky decomposition. If $M' \notin \mathcal{S}_d^{++}$, then we know by Theorem 3.2.2 that it does not belong to $\mathcal{M}_d$. If $M' \in \mathcal{S}_d^{++}$, then by Lemma 3.2.7, it suffices to check whether $\langle e, Me\rangle \geq 1$ for the finitely many $e \in \mathcal{Z}_d$ satisfying $|e|^2 \leq \lambda_{\min}(M)^{-1}$.

At step 5, we know that M_0 has to have the same determinant as M', therefore, at least in dimension $d \leq 4$, there is at most one suitable candidate $M_0 \in \mathrm{Perfect}_0(d)$ (and we observe that there is always exactly one). It remains to find $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $M' = A^\top M_0 A$. To this end, we first compute $\Xi(M_0)$ and $\Xi(M')$. Then it would be natural to look for $A \in \mathrm{GL}_d(\mathbb{Z})$ such that $\Xi(M') = \{A^{-1}e \mid e \in \Xi(M_0)\}$, following Proposition 3.2.4. Since we observe that finding A is the more numerically costly part of our implementation of Voronoi's algorithm, we use the following lemma for improved efficiency:

Lemma 3.A.1. *Let M, $M' \in \mathrm{Perfect}(d)$. Assume that $\det(M) = \det(M')$, that $\Xi(M)$ and $\Xi(M')$ have the same cardinality $I \geq d$, and that one may index elements of $\Xi(M)$ and $\Xi(M')$:*

$$\Xi(M) = \{e_i \mid 1 \leq i \leq I\}, \qquad \Xi(M') = \{e'_i \mid 1 \leq i \leq I\},$$

such that $\det(e_1, \ldots, e_d) = \pm 1$ and $\langle e_i, Me_j\rangle = \langle e'_i, M'e'_j\rangle$, for any $1 \leq i \leq j \leq I$. Let B (respectively B') be the matrix whose columns are $(e_1, \ldots, e_d)$ (respectively $(e'_1, \ldots, e'_2)$), and let $A := B'B^{-1}$. Then $A \in \mathrm{GL}_d(\mathbb{Z})$ and $M = A^\top M' A$.

Proof. One has $B^\top MB = (B')^\top M'B'$, and thus $M = A^\top M'A$. Moreover $B \in \mathrm{GL}_d(\mathbb{Z})$, thus $B^{-1} \in \mathrm{GL}_d(\mathbb{Z})$ and A has integer components, as a product of matrices with integer components. Since $\det(M) = \det(M')$, one has $\det(A) = \pm 1$, therefore $A \in \mathrm{GL}_d(\mathbb{Z})$. □

We observe that there always exist indexings of $\Xi(M_0)$ and $\Xi(M')$ satisfying the assumptions of Lemma 3.A.1. Finding those indexings reduces to finding an isomorphism between the graphs whose nodes are elements of $\Xi(M_0)$ (respectively $\Xi(M')$) and whose edge between two nodes e_1, e_2 in $\Xi(M_0)$ (respectively $\Xi(M')$) exists and has label $\langle e_1, M_0 e_2\rangle$ (respectively $\langle e_1, M'e_2\rangle$). This may be achieved efficiently using well-known algorithms such as [Cor+04].

Part II

Monotone discretization of some specific degenerate elliptic partial differential equations

Chapter 4

A linear finite-difference scheme for approximating Randers distances on Cartesian grids

This chapter corresponds to the paper [BBM21a].

4.1 Introduction

A Randers metric is the sum of a Riemannian metric and of an anti-symmetric perturbation, suitably bounded and defined by linear form. By construction, a Randers metric is in general non-symmetric, and so is the associated path-length distance, see Remark 4.1.3 on terminology. Such metrics can account, in a very simple manner, for the fact that moving a vehicle uphill, or advancing a boat against water currents, costs more than the opposite operation. The asymmetry embedded in Randers metrics opens up numerous applications which cannot be addressed with the simpler Riemannian metrics, ranging from general relativity [Ran41] to image segmentation [CMC16], through quantum vortices [ABM06] and path curvature penalization [CMC17], see Remark 4.1.1.

In this paper, we present a numerical scheme for computing Randers distances by solving a *linear second order* Partial Differential Equation (PDE). Our approach is based on a generalization of Varadhan's formula [Var67], which is commonly used to compute Riemannian distances [CWW13]. Let us emphasize that Randers distances also obey a *non-linear first order* PDE [BC97], which can be solved directly numerically [Mir14; Mir19], yet the linear structure of the PDE formulation considered in this paper has a number of computational advantages, including easier numerical implementation, faster computation in some cases, and smoothness of the numerical solution, see Remark 4.1.2. Some of our results, such as the identification of the optimal scaling of the relaxation parameter ε w.r.t. the grid scale h, and the proof of convergence in the case of point sources, are new as well in the special cases of isotropic and Riemannian metrics. We present an application to numerical optimal transportation, enabled by the linear structure of the discretization, with an asymmetric cost function defined as the Randers distance between the source and target, generalizing previous works limited to Riemannian costs [Cut13].

In order to make our statements more precise, we need to introduce some notations. Throughout this paper, $\Omega \subset \mathbb{R}^d$ denotes a smooth bounded and connected open domain, equipped with a metric $\mathcal{F} : \overline{\Omega} \times \mathbb{R}^d \to [0, \infty[$, whose explicit form is discussed below (4.2). The corresponding

path-length and distance are defined by

$$\text{length}_{\mathcal{F}}(\gamma) := \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t))\,\mathrm{d}t, \qquad \text{dist}_{\mathcal{F}}(x,y) := \inf_{\gamma\in\Gamma_x^y} \text{length}_{\mathcal{F}}(\gamma). \tag{4.1}$$

We denoted by γ an element of the collection $\Gamma := \text{Lip}([0,1],\overline{\Omega})$ of locally Lipschitz paths within the domain closure, and by $\Gamma_x^y \subset \Gamma$ the subset of paths from $x \in \overline{\Omega}$ to $y \in \overline{\Omega}$. We assume in this paper that $\mathcal{F}$ has the structure of a Randers metric: there exists a field $M : \overline{\Omega} \to S_d^{++}$ of symmetric positive definite matrices, and a vector field $\omega : \overline{\Omega} \to \mathbb{R}^d$, both having Lipschitz regularity, and such that for all $x \in \overline{\Omega}$ and all $v \in \mathbb{R}^d$ one has

$$\mathcal{F}_x(v) := |v|_{M(x)} + \langle \omega(x), v\rangle, \qquad \text{where } |\omega(x)|_{M(x)^{-1}} < 1. \tag{4.2}$$

We denoted by $\langle\cdot,\cdot\rangle$ the standard Euclidean scalar product, and by $|v|_M := \sqrt{\langle v, Mv\rangle}$ the anisotropic (but symmetric) norm on $\mathbb{R}^d$ defined by a symmetric positive definite matrix M. The smallness constraint (4.2, right) ensures that $\mathcal{F}_x(v) > 0$ for all $v \neq 0$, $x \in \overline{\Omega}$. Randers metrics include Riemannian metrics as a special case, when the vector field ω vanishes identically over the domain. See Figure 4.2 for an illustration of their unit balls, distance maps, and minimal paths.

Our approach to the computation of Randers distances goes through the solution to a linear second order PDE, depending on a small parameter $\varepsilon > 0$, and some boundary data $g \in C^0(\partial\Omega, \mathbb{R})$

$$u_\varepsilon + 2\varepsilon\langle b, \nabla u_\varepsilon\rangle - \varepsilon^2 \,\text{Tr}(A_b \nabla^2 u_\varepsilon) = 0 \text{ in } \Omega, \qquad u_\varepsilon = \exp(-g/\varepsilon) \text{ on } \partial\Omega, \tag{4.3}$$

where A_b is a field of symmetric positive definite matrices, and b is a vector field, depending in a simple algebraic manner on the Randers metric parameters M and ω, see Lemma 4.2.6 and (4.9). In the Riemannian special case one has $A_b = M^{-1}$ and $b = \omega = 0$, consistently with [Var67]. We establish in Theorem 4.2.12, following [BP88], that for all $x \in \Omega$

$$\mathbf{u}(x) := \lim_{\varepsilon\to 0} -\varepsilon \ln u_\varepsilon(x) \qquad \text{exists and satisfies} \quad \mathbf{u}(x) = \min_{p\in\partial\Omega} g(p) + \text{dist}_{\mathcal{F}}(p,x). \tag{4.4}$$

In other words, $\mathbf{u}$ is the Randers distance from the boundary $\partial\Omega$, with initial time penalty g, see section 4.4 for the case of point sources. Note that one often considers the opposite problem, of reaching a boundary point $p \in \partial\Omega$ from x, which is equivalent up to replacing the vector field ω with its opposite in (4.2), see Definition 4.2.3 and the discussion below. The distance map $\mathbf{u}$ also obeys the first order non-linear Hamilton-Jacobi-Bellman equation

$$|\nabla \mathbf{u} - \omega|_{M^{-1}} = 1 \text{ in } \Omega, \qquad \mathbf{u} = g \text{ on } \partial\Omega, \tag{4.5}$$

in the sense of viscosity solutions (possibly with discontinuous boundary conditions) [BC97], which is numerically tractable [Mir14; Mir19] as well. The point of this paper is however to study the linear approach (4.3) which has a number of advantages, see Remark 4.1.2. We present a finite difference discretization of (4.3) on the Cartesian grid $\Omega_h := \Omega \cap h\mathbb{Z}^d$, of dimension $d \in \{2,3\}$, denoting by $h > 0$ the grid scale, reading

$$u + 2\varepsilon \sum_{1\le i\le I} \rho_i \langle A_b^{-1} b, e_i\rangle \,\overline{\delta}_h^{e_i} u - \varepsilon^2 \sum_{1\le i\le I} \rho_i \Delta_h^{e_i} u = 0 \quad \text{ on } \Omega_h, \tag{4.6}$$

where $\overline{\delta}_h^e$ and Δ_h^e denote standard centered and second order finite differences (4.26), modified close to $\partial\Omega$ to account for the Dirichlet boundary conditions, see (4.27) and (4.28). We denoted by $\rho_i(x) \geq 0$ and $e_i(x) \in \mathbb{Z}^d$, $1 \leq i \leq I = d(d+1)/2$ the weights and offsets of Selling's decomposition [Sel74; CS92] of the matrix $A_b(x)$, a tool from lattice geometry which is convenient for the design

of anisotropic finite differences schemes [FM14; Mir18; Mir19] in dimension $d \in \{2,3\}$, see section 4.B. Denoting by u_ε^h the solution of (4.6) we prove in Theorem 4.3.18 that $-\varepsilon \ln u_\varepsilon^h \to \mathbf{u}$ as $(\varepsilon, h/\varepsilon) \to 0$. The case of point sources also requires $\varepsilon \ln h \to 0$, see Theorem 4.4.1. The optimal consistency order is achieved when $\varepsilon = h^{\frac{2}{3}}$, see Corollary 4.3.14.

Finally we present in section 4.5 an application to the optimal transport problem

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\,\mathrm{d}\gamma(x,y), \qquad \text{with } c(x,y) := \mathrm{dist}_{\mathcal{F}}(x,y), \tag{4.7}$$

where μ and ν are given probability measures on Ω, and $\Pi(\mu,\nu)$ is the set of probability measures on $\Omega \times \Omega$ whose first and second marginals coincide respectively with μ and ν. The proposed implementation relies on Sinkhorn's matrix scaling algorithm [Sin64], and the linear structure of (4.3). We emphasize that the matrix of costs $(c(x,y))_{x,y\in\Omega_h}$ is never numerically constructed, and would not fit in computer memory, but instead that the adequate matrix-vector product are evaluated by solving finite difference equations similar to (4.6), in an efficient manner thanks to a preliminary sparse matrix factorization. Let us acknowledge here that, in contrast with the Riemannian case [Sol+15], our approach does not extend to the quadratic cost $c(x,y) = \mathrm{dist}_{\mathcal{F}}(x,y)^2$. Indeed, this specific cost is handled in the Riemannian case using the short time asymptotic estimates of the diffusion equation, which becomes non-linear in the case of Randers geometry, see Remark 4.4.5, in contrast with the Poisson equation (4.3).

Contributions We establish that the solution to a linear second order PDE converges, as a relaxation parameter $\varepsilon \to 0$ and after a logarithmic transformation, to the Randers distance from the domain boundary. We propose a finite difference discretization of that linear PDE, on a Cartesian grid of scale h, and establish convergence of the numerical solutions as $\varepsilon \to 0$ and $h/\varepsilon \to 0$, with optimal consistency when $\varepsilon = h^{\frac{2}{3}}$. We extend the approach to the case of point sources, under the additional condition $\varepsilon \ln h \to 0$. We propose a computational method for optimal transport with Randers distance as cost. Numerical experiments illustrate our results.

Outline We recall in section 4.2 the definition of Randers distances and introduce an extension of Varadhan's formula to Randers manifolds. We describe the coefficients of (4.3) in terms of the Randers metric (4.2), and prove the convergence result (4.4).

We study in section 4.3 the linear finite-difference scheme (4.6). We show that a logarithmic transformation of the solution (4.6) solves another nonlinear scheme, for which we prove convergence and consistency with the non-linear PDE (4.5). We also discuss heuristic techniques introduced in [CWW13] to improve the numerical accuracy of the geodesic distance approximation, and extend them to Randers metrics.

We address in section 4.4 the computation of the geodesic distance from a point source, and in section 4.5 the discretization of the optimal transportation problem (4.7), extending [Sol+15] which is limited to Riemannian distance costs.

Finally, we illustrate in section 4.6 our results with numerical experiments, devoted to the computation of Randers distances and of the corresponding geodesic paths, and to the solution of the optimal transport problem (4.7) on a Randers manifold.

Remark 4.1.1 (Applications of Randers metrics). Randers metrics are, arguably, the simplest model of a *non-symmetric metric*, often referred to as a quasi-metric, see Remark 4.1.3. They play a key role in Zermelo's problem [BRS04] of path planning subject to a drift, see section 4.2.2, but also have numerous independent applications, of which we can only give a glimpse here. The common feature of these applications is that the paths are naturally endowed with an orientation.

The boundary of a simply connected image region, oriented counterclockwise, minimizes the classical Chan-Vese segmentation functional iff it is a minimal geodesic for a suitable Randers metric, leading to a robust numerical optimization method [CMC16]. The Euler-Mumford elastica minimal path model, whose cost is defined by integrating the squared curvature (plus a constant), is a limiting case of a Randers model, which allows the numerical computation of global minimizers with applications to tubular structure extraction in images [CMC17]. Quantum vortex filaments, in a suitable limit and under appropriate assumptions, follow Randers geodesics, where the asymmetric part of the metric is derived from the magnetic field [ABM06]. Finally, let us mention that Randers metrics were introduced in the context of general relativity, where the trajectory orientation stems from the time coordinate induced by the Minkowski space-time quadratic form [Ran41].

Remark 4.1.2 (Advantages of linear schemes for distance map computation). Distance maps are ubiquitous in mathematics and their applications, and a variety of approaches have been proposed for their numerical computation [Cra+20], including Randers distances [Mir14; Mir19]. The use of a linear PDE (4.3), is here largely motivated by its application to the optimal transport problem (4.7), but this approach has other advantages, see [CWW13] for a more detailed discussion:

- (Ease of implementation) While we limit here our attention to domains discretized on Cartesian grids, geodesic distance computation also makes sense on manifolds presented as triangulations [CWW13], patch based surfaces, etc. In that context, discretizing the non-linear PDE (4.5) can be challenging, whereas standard tools are often available for linear PDEs such as (4.6).

- (Computation speed) Factorization techniques for sparse linear systems of equations are a subject of continued research, including non-symmetric Laplacian-like operators [Coh+18]. Once the linear system (4.6) is factored, it can be solved for a modest cost with a large number of right-hand sides, for instance to compute all pairwise Randers distances within a set of points, or when addressing the optimal transport problem (4.7) using Sinkhorn's matrix scaling algorithm as described in section 4.5.

- (Solution smoothness) The distance map $\mathbf{u}$ defined by (4.4) is non-differentiable across the cut-locus[1], and numerical solvers [Mir14; Mir19] of the generalized eikonal PDE (4.5) produce non-smooth approximations of it. In contrast, the solution to the linear equation (4.3) is smooth and yields a natural regularization $\mathbf{u}_\varepsilon := -\varepsilon \ln u_\varepsilon$, for any $\varepsilon > 0$, of the limit distance map $\mathbf{u}$.

Remark 4.1.3 (*quasi-* prefix and asymmetric geometry). Non-symmetric norms, metrics and distances are often referred to as *quasi*-norms, *quasi*-metrics and *quasi*-distances. However, we drop the prefix "quasi" in this paper for the sake of readability and uniformity.

Conventions and notations We denote by $|\cdot|$ the Euclidean norm on $\mathbb{R}^d$, and by $\mathcal{S}_d$, $\mathcal{S}_d^+$, and $\mathcal{S}_d^{++}$ the sets of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices of size $d \times d$ respectively. For any $A, B \in \mathcal{S}_d$, the relation $A \succeq B$ stands for $A - B \in \mathcal{S}_d^+$ (resp. $A \succ B$ stands for $A - B \in \mathcal{S}_d^{++}$), which is the Loewner order on symmetric matrices. For any $A \in \mathcal{S}_d^{++}$ and $b \in \mathbb{R}^d$, we define

$$\|A\| := \sup_{|x| \leq 1} |Ax|, \qquad\qquad |b|_A := \langle b, Ab \rangle^{1/2}.$$

[1]The cut locus is the set of points where the minimum (4.4, right) is attained by several minimal paths from the boundary.

From now on, we consider an open, bounded, connected, and nonempty domain $\Omega \subset \mathbb{R}^d$ with a $W^{3,\infty}$ boundary. The unknowns to the partial differential equations, and to their discretization schemes, are distinguished by typography: u for the linear second order PDEs (4.3) or numerical scheme (4.6) and variants, and $\mathbf{u}$ for the non-linear PDE (4.5) and related.

4.2 Elements of Randers geometry

A Randers metric is defined as the sum of a Riemannian metric, and of a suitably bounded linear term (4.2). We present section 4.2.1 these geometrical objects in more detail, making connections with Zermelo's navigation problem section 4.2.2. The eikonal equation (4.5) is discussed in section 4.2.3, and its linear variant (4.3) in section 4.2.4. We establish in Theorem 4.2.12 the existence of a solution u_ε to the linear PDE (4.3), and the convergence of $\mathbf{u}_\varepsilon = -\varepsilon \ln u_\varepsilon$ to the value function of the arrival time problem (4.4) as the relaxation parameter $\varepsilon > 0$ vanishes. The proof, based on the theory of viscosity solutions to degenerate elliptic PDEs, is postponed to section 4.A.

Before specializing to the case of Randers geometry, we briefly recall here the generic or *Finslerian* definition of a *non-symmetric* norm, dual-norm, metric, and path-length distance, and some of their elementary properties.

Definition 4.2.1. A function $F : \mathbb{R}^d \to \mathbb{R}_+$ is a norm iff it is convex, 1-homogeneous, and vanishes only at the origin. The dual norm $F^* : \mathbb{R}^d \to \mathbb{R}_+$ is defined for all $v \in \mathbb{R}^d$ by

$$F^*(v) := \max_{F(w)\leq 1} \langle v, w\rangle. \tag{4.8}$$

Equivalently, by homogeneity of F, one has $F^*(v) := \max\{\langle v, w\rangle / F(w);\ |w| = 1\}$. Conventionally, the above defines a *quasi*-norm, whereas a norm is subject to the additional symmetry axiom $F(v) = F(-v)$ for all $v \in \mathbb{R}^d$. However the prefix "quasi" before norms, metrics and distances is dropped in this paper for readability, as already mentioned in Remark 4.1.3. The following facts, stated without proof, are standard results of convex analysis and Finsler geometry.

Lemma 4.2.2 (Norm duality)**.** *Any norm F on $\mathbb{R}^d$ is Lipschitz continuous on $\mathbb{R}^d$, and as a result the extremum in* (4.8) *is indeed attained, for any $w \in \mathbb{R}^d$. The dual norm F^* is also a norm, and furthermore one has $F^{**} = F$ identically on $\mathbb{R}^d$.*

Definition 4.2.3. A metric on a domain $\Omega \subset \mathbb{R}^d$ is a continuous map $\mathcal{F} : \overline{\Omega} \times \mathbb{R}^d \to \mathbb{R}_+$, denoted $(x, v) \mapsto \mathcal{F}_x(v)$, such that $\mathcal{F}_x$ is a norm on $\mathbb{R}^d$ for all $x \in \overline{\Omega}$. The dual metric $\mathcal{F}^*$ is defined pointwise from the dual norms. The related path length and distance are defined from (4.1) and denoted $\mathrm{length}_{\mathcal{F}}$ and $\mathrm{dist}_{\mathcal{F}}$.

Let us emphasize that $\mathrm{dist}_{\mathcal{F}}(x, y) \neq \mathrm{dist}_{\mathcal{F}}(y, x)$ in general, for $x, y \in \overline{\Omega}$, since norms and metrics are not assumed here to be symmetric. In the special case where $\mathcal{F}_x = F$ is a constant metric, and $[x, y] \subset \overline{\Omega}$, one has $\mathrm{dist}_{\mathcal{F}}(x, y) = F(y - x)$.

Lemma 4.2.4 (Path-length distance)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded connected domain with smooth boundary and equipped with a metric $\mathcal{F}$. Then the extremum* (4.1) *defining* $\mathrm{dist}_{\mathcal{F}}(x, y)$ *is attained, for any $x, y \in \overline{\Omega}$, and defines a distance over $\overline{\Omega}$. Furthermore there exists $0 < c \leq C$ such that $c|x - y| \leq \mathrm{dist}_{\mathcal{F}}(x, y) \leq C|x - y|$ for all $x, y \in \overline{\Omega}$.*

4.2.1 Algebraic structure of Randers metrics

Randers norms are defined by analogy to Randers metrics (4.2), as the sum of a symmetric part defined from a symmetric positive definite matrix, and of an anti-symmetric linear part.

Definition 4.2.5. A Randers norm $F : \mathbb{R}^d \to \mathbb{R}^d$ takes the form $F(v) = |v|_M + \langle \omega, v \rangle$, where $M \in \mathcal{S}_d^{++}$, and $\omega \in \mathbb{R}^d$ is subject to $|\omega|_{M^{-1}} < 1$.

The dual to a Randers norm also is a Randers norm, as shown by the following lemma, whose proof can be found in Proposition 4.1 of [Mir14] and Appendix C of [Mir19].

Lemma 4.2.6 (Randers dual norm [Mir14; Mir19])**.** *The dual to a Randers norm F of parameters M, ω is also a Randers norm, of parameters A, b characterized by the following relation, where $\alpha > 0$:*

$$\begin{pmatrix} A/\alpha & b \\ b^\top & \alpha \end{pmatrix} = \begin{pmatrix} M & \omega \\ \omega^\top & 1 \end{pmatrix}^{-1}.$$

Note that α in Lemma 4.2.6 is determined as the bottom right coefficient in $\begin{pmatrix} M & \omega \\ \omega^\top & 1 \end{pmatrix}^{-1}$. In the special case where $\omega = 0$, one obtains $A = M^{-1}$, $b = 0$, and $\alpha = 1$, recovering the well known fact that the dual to a Riemannian norm is also a Riemannian norm, defined by the inverse symmetric matrix. The duality formula in Lemma 4.2.6 is only the first of a family of algebraic identities associated with Randers norms, presented in Lemma 4.2.7 below, and used to reformulate the PDEs (4.3) and (4.5). For that purpose, we need to introduce some notation. For any $A \in \mathcal{S}_d$, and any $b \in \mathbb{R}^d$ we let

$$A_b := A - bb^\top. \tag{4.9}$$

The Schur complement formula yields the following positive-definiteness equivalences:

$$A_b \succ 0 \quad \Leftrightarrow \quad \begin{pmatrix} A & b \\ b^\top & 1 \end{pmatrix} \succ 0 \quad \Leftrightarrow \quad (A \succ 0 \text{ and } |b|_{A^{-1}} < 1). \tag{4.10}$$

If $M : \overline{\Omega} \to S_d^{++}$ and $\omega : \overline{\Omega} \to \mathbb{R}^d$ are Lipschitz fields obeying $|\omega|_{M^{-1}} < 1$ pointwise on $\overline{\Omega}$ (which is compact by assumption), then the fields $A : \overline{\Omega} \to S_d^{++}$ and $b : \overline{\Omega} \to \mathbb{R}^d$ defined by Lemma 4.2.6 as the dual Randers parameters are also Lipschitz, since matrix inversion is differentiable, and obey the equivalent properties (4.10) pointwise on $\overline{\Omega}$, thus $|b|_{A^{-1}} < 1$. The following lemma provides several equivalent characterizations of the unit ball associated with a Randers norm, and ends this subsection.

Lemma 4.2.7. *Let M, ω denote the parameters of a Randers norm, and A, b the parameters of the dual Randers norm, see Lemma 4.2.6. Then for all $v \in \mathbb{R}^d$*

$$\Big[|v|_M + \langle \omega, v \rangle - 1\Big] \quad \propto \quad \Big[|v|^2_{M_\omega} + 2\langle \omega, v \rangle - 1\Big] \quad \propto \quad \Big[|v - b|_{A^{-1}} - 1\Big] \tag{4.11}$$

$$\Big[|v|_A + \langle b, v \rangle - 1\Big] \quad \propto \quad \Big[|v|^2_{A_b} + 2\langle b, v \rangle - 1\Big] \quad \propto \quad \Big[|v - \omega|_{M^{-1}} - 1\Big], \tag{4.12}$$

where $x \propto y$ means that $\operatorname{sign}(x) = \operatorname{sign}(y)$, with $\operatorname{sign} : \mathbb{R} \to \{-1, 0, 1\}$ the sign function.

Proof. Note that the second line can be deduced from the first one, by exchanging the role of the Randers norm and of its dual norm. The positive definiteness of A_b and M_ω follows from (4.10) and Definition 4.2.5. Under the assumptions of the lemma, one has the equivalences

$$\begin{aligned} |v|_M + \langle \omega, v \rangle - 1 \leq 0 \;&\Leftrightarrow\; |v|_M \leq 1 - \langle \omega, v \rangle \;\Leftrightarrow\; |v|^2_M \leq (1 - \langle \omega, v \rangle)^2 \\ &\Leftrightarrow\; |v|_{M_\omega} + 2\langle \omega, v \rangle - 1 \leq 0, \end{aligned}$$

and likewise with strict inequalities, which implies (4.11, left equivalence). The only difficulty lies in the reverse implication of the second equivalence: we must exclude the case where

$|v|_M \leq \langle \omega, v\rangle - 1$, and indeed this is in contradiction with $|\langle \omega, v\rangle| \leq |\omega|_{M^{-1}}|v|_M < |v|_M + 1$ since $|\omega|_{M^{-1}} < 1$ by assumption.

Denoting by F the Randers norm of parameters M, ω, and by F^* the dual norm, one has

$$|v-b|_{A^{-1}} \leq 1 \Leftrightarrow \big(\forall w, \langle w, v-b\rangle \leq |w|_A\big) \Leftrightarrow \big(\forall w, \langle w, v\rangle \leq |w|_A + \langle b, w\rangle := F^*(w)\big) \Leftrightarrow F(v) \leq 1,$$

where implicitly $w \in \mathbb{R}^d$. In the last equivalence we used $F(v) = F^{**}(v) = \max\{\langle v, w\rangle; F^*(w) \leq 1\}$. A similar equivalence can be obtained with strict inequalities for any $w \neq 0$, which concludes the proof of (4.11, right equivalence) and of this lemma. □

4.2.2 Zermelo's navigation problem

Zermelo [BRS04] considers a vehicle able to move at speed at most $c(x)$ *relative* to a given medium, which itself is subject to a drift $\eta(x)$, where $x \in \overline{\Omega}$ is the position. Typically, the vehicle is described as a boat subject to water currents, or as a flying object subject to air currents.

The set admissible *absolute* velocities v at the point x is characterized by the following relation

$$|v - \eta(x)| \leq c(x). \tag{4.13}$$

Given two endpoints $x, y \in \overline{\Omega}$, Zermelo's navigation problem asks for the smallest time $T = T_c^\eta(x,y) \geq 0$ such that there exists $\gamma \in \mathrm{Lip}([0,T], \overline{\Omega})$ obeying $|\gamma'(t) - \eta(\gamma(t))| \leq c(\gamma(t))$ for a.e. $t \in [0,T]$, and $\gamma(0) = x$, $\gamma(T) = y$. In other words, $T_c^\eta(x,y)$ is the minimal time from x to y subject to the velocity constraints (4.13).

The vehicle described by Zermelo's problem is locally controllable at $x \in \Omega$ iff $|\eta(x)| < c(x)$, in other words iff the drift velocity norm is smaller than the maximum relative vehicle speed. In that case, it can be reformulated as a Randers minimal path problem.

Proposition 4.2.8. *Let $c : \overline{\Omega} \to \mathbb{R}$, and $\eta : \overline{\Omega} \to \mathbb{R}^d$ be continuous and obey $c > 0$ and $|\eta| < c$, pointwise on $\overline{\Omega}$. Consider the Randers metric $\mathcal{F}^*$ of parameters $A = c^2 I_d$ and $b = \eta$ on Ω, as well as its dual $\mathcal{F}^{**} = \mathcal{F}$. Then for all $x, y \in \overline{\Omega}$*

$$T_c^\eta(x,y) = \mathrm{dist}_{\mathcal{F}}(x,y).$$

Proof. Let $M : \overline{\Omega} \to S_d^{++}$ and $\omega : \overline{\Omega} \to \mathbb{R}^d$ be parameters of the Randers metric $\mathcal{F}$. The distance $\mathrm{dist}_{\mathcal{F}}(x,y)$ is the smallest time T for which there exists a path $\gamma \in \mathrm{Lip}([0,T], \overline{\Omega})$ obeying $1 \geq \mathcal{F}_{\gamma(t)}(\gamma'(t)) := |\gamma'(t)|_{M(\gamma(t))} + \langle \omega(\gamma(t)), \gamma'(t)\rangle$ for a.e. $t \in [0,T]$, and $\gamma(0) = x$, $\gamma(T) = y$. Indeed, this follows from the definition (4.1) and by reparametrization of any Lipschitz path at unit speed w.r.t. the metric $\mathcal{F}$. From this point, the announced result follows from the equivalence of $1 \geq \mathcal{F}_x(v) := |v|_{M(x)} + \langle \omega(x), v\rangle$ with (4.13), established in (4.11). □

4.2.3 The Eikonal equation

Consider a domain Ω, equipped with a Randers metric $\mathcal{F}$ with Lipschitz coefficients on Ω, and penalty function $g \in C^0(\partial\Omega, \mathbb{R})$. We are interested in the following value function $\mathbf{u} : \overline{\Omega} \to \mathbb{R}$, corresponding to the minimal time to reach $x \in \overline{\Omega}$ from a boundary point $p \in \partial\Omega$, with initial time penalty $g(p)$:

$$\mathbf{u}(x) := \min_{p \in \partial\Omega} g(p) + \mathrm{dist}_{\mathcal{F}}(p,x). \tag{4.14}$$

We prove in Theorem 4.A.9 that (4.14) is a viscosity solution, see Definition 4.A.2, to the first order non-linear PDE

$$\mathcal{F}_x^*(\nabla \mathbf{u}(x)) = 1 \text{ for all } x \in \Omega, \qquad \mathbf{u}(x) = g(x) \text{ for all } x \in \partial\Omega. \tag{4.15}$$

The boundary condition $\mathbf{u} = g$ on $\partial\Omega$ is satisfied in a strong sense if $g(x) \leq g(p) + \mathrm{dist}_{\mathcal{F}}(p, x)$ for all $x, p \in \partial\Omega$, but in the weak sense of Definition 4.A.2 otherwise. The comparison principle Theorem 4.A.8 implies that the viscosity solution is uniquely determined in Ω.

Corollary 4.2.9. *If $\mathcal{F}$ is a Randers metric of parameters M, ω, and dual parameters A, b, then the eikonal PDE (4.15, left) admits the following three equivalent formulations in Ω in the sense of viscosity solutions*

$$|\nabla \mathbf{u}|_A + \langle \nabla \mathbf{u}, b\rangle = 1, \qquad |\nabla \mathbf{u}|^2_{A_b} + 2\langle \nabla \mathbf{u}, b\rangle = 1, \qquad |\nabla \mathbf{u} - \omega|_{M^{-1}} = 1. \tag{4.16}$$

Proof. The equation $\mathcal{F}^*_x(\nabla \mathbf{u}(x)) = 1$ is a shorthand for (4.16, left) at $x \in \Omega$, see Definition 4.2.5 of a Randers norm. It is equivalent to (4.16, center and right) by (4.12). □

In applications, computing the value function (4.14) is often only a means to obtain the globally optimal path γ from $\partial\Omega$ to an arbitrary point $x_* \in \Omega$. This path can be extracted by solving, backwards in time, the following Ordinary Differential Equation (ODE), see e.g. [Dui+18, Appendix C]

$$\gamma'(t) = V(\gamma(t)), \qquad \text{where } V(x) := \mathrm{d}\mathcal{F}^*_x(\nabla \mathbf{u}(x)) \tag{4.17}$$

for all $x \in \overline{\Omega}$. The ODE needs to be solved on the interval $[0, T]$ where $T = \mathbf{u}(x_*)$, with terminal condition $\gamma(T) = x_*$. By $\mathrm{d}\mathcal{F}^*_x(w)$ we denote the derivative of $\mathcal{F}^*_x$ w.r.t. the variable w, where $x \in \overline{\Omega}$ is fixed.

Corollary 4.2.10. *The following expressions are positively proportional to the geodesic flow V defined by (4.17, right), at all points where $\mathbf{u}$ is differentiable*

$$\frac{A\nabla \mathbf{u}}{|\nabla \mathbf{u}|_A} + b, \qquad A_b \nabla \mathbf{u} + b, \qquad M^{-1}(\nabla \mathbf{u} - \omega). \tag{4.18}$$

Proof. Fix a point $x \in \Omega$ where $\mathbf{u}$ is differentiable, and denote $v := \nabla \mathbf{u}(x)$. Introduce the Randers norm $F^* = \mathcal{F}^*_x$ whose parameters are denoted $A \in S^{++}_d$ and $b \in \mathbb{R}^d$, in such way that $F^*(v) = 1$ by (4.15). Differentiating $F^*(v) = |v|_A + \langle b, v\rangle$ we obtain $\mathrm{d}F(v) = Av/|v|_A + b$ which yields (4.18, left). The three expressions (4.12) vanish, and their respective gradients w.r.t. v are $g_1 := Av/|v|_A + b$, $g_2 := 2(A_b v + b)$ and $g_3 := M^{-1}(v - \omega)/|v - \omega|_{M^{-1}}$. These gradients are non-zero since $\langle v, g_1\rangle = F^*(v) = 1$, $\langle v, g_2\rangle = 1 + |v|^2_{A_b} \geq 1$ and $\langle v - \omega, g_3\rangle = |v - \omega|^2_{M^{-1}} = 1$. Since g_1, g_2 and g_3 are orthogonal to the same level set, and point outward of it, they are positively proportional. The result follows. □

4.2.4 Varadhan's formula

Varadhan's formula is based on a logarithmic transformation of the unknown [Var67], which turns the linear PDE (4.19) into the non-linear PDE (4.20). The point of this transformation is that, with a proper scaling of the unknown and the PDE coefficients, a relaxation parameter $\varepsilon > 0$ is eliminated from the boundary conditions and from all the PDE coefficients except one, of principal order.

Lemma 4.2.11 (Logarithmic transformation)**.** *Let $\varepsilon > 0$, and let u_ε be a viscosity solution to*

$$u + 2\varepsilon\langle \nabla u, b\rangle - \varepsilon^2 \operatorname{Tr}(A_b \nabla^2 u) = 0 \text{ in } \Omega, \qquad u = \exp(-g/\varepsilon) \text{ on } \partial\Omega, \tag{4.19}$$

where $\Omega \subset \mathbb{R}^d$ is a smooth bounded domain, $A_b : \overline{\Omega} \to S^{++}_d$ and $b : \overline{\Omega} \to \mathbb{R}^d$ are Lipschitz, and $\varepsilon > 0$. Then $\mathbf{u}_\varepsilon := -\varepsilon \ln u_\varepsilon$ is a viscosity solution to the PDE

$$|\nabla \mathbf{u}|^2_{A_b} + 2\langle \nabla \mathbf{u}, b\rangle - \varepsilon \operatorname{Tr}(A_b \nabla^2 \mathbf{u}) = 1 \text{ in } \Omega, \qquad \mathbf{u} = g \text{ on } \partial\Omega. \tag{4.20}$$

Lemma 4.2.11 is an immediate consequence of Corollary 4.A.5 established in section 4.A. For later convenience, we introduce the following PDE operators on the domain Ω

$$\mathcal{L}^\varepsilon u = u + 2\varepsilon\langle\nabla u, b\rangle - \varepsilon^2 \operatorname{Tr}(A_b\nabla^2 u), \quad \mathcal{S}^\varepsilon \mathbf{u} = |\nabla \mathbf{u}|^2_{A_b} + 2\langle\nabla\mathbf{u}, b\rangle - \varepsilon \operatorname{Tr}(A_b\nabla^2\mathbf{u}) - 1, \quad (4.21)$$

and observe that, formally, one has $\mathcal{S}^\varepsilon \mathbf{u} = -e^{\frac{\mathbf{u}}{\varepsilon}}\mathcal{L}^\varepsilon(e^{-\frac{\mathbf{u}}{\varepsilon}})$. The following result relies on the framework of viscosity solutions to take the limit $\varepsilon \to 0$ in $\mathcal{S}^\varepsilon$, letting the second order "viscous" term $-\varepsilon \operatorname{Tr}(A_b\nabla^2\mathbf{u})$ vanish, and recovering in the limit a first order non-linear equation equivalent to the Randers eikonal equation, see Corollary 4.2.9.

Theorem 4.2.12 (Vanishing viscosity limit)**.** *The PDE* (4.19) *admits a unique viscosity solution in* Ω*. In addition* $\mathbf{u}_\varepsilon := -\varepsilon \ln u_\varepsilon$ *converges locally uniformly in* Ω *to the value function* (4.14)*, associated with the Randers metric* $\mathcal{F}$ *whose dual metric* $\mathcal{F}^*$ *has parameters* A, b*.*

The elements of proof relying on the concept of viscosity solutions are postponed to section 4.A. In particular, uniqueness of the solution to (4.19) follows from the comparison principle Proposition 4.A.7, see [BP88]. Convergence as $\varepsilon \to 0$ is established in Theorem 4.A.12. We limit our attention here to the existence of a solution to (4.19), which is based on the interpretation of u_ε as an expectation of a cost associated with a stochastic process. Fix $\varepsilon > 0$, and introduce the stochastic process $(X_t^{x,\varepsilon})_{t\geq 0}$

$$dX_t^{x,\varepsilon} = -2\varepsilon b(X_t^{x,\varepsilon})\,dt + \varepsilon\sqrt{2A_b}(X_t^{x,\varepsilon})\,dW_t, \qquad X_0^{x,\varepsilon} = x, \quad (4.22)$$

where $(W_t)_{t\geq 0}$ is a d-dimensional Wiener process. Define also the exit time $\tau^{x,\varepsilon}$ by

$$\tau^{x,\varepsilon} := \inf\{t \geq 0;\ X_t^{x,\varepsilon} \notin \Omega\}.$$

Since Ω is bounded, and A_b is positive definite, the exit time $\tau^{x,\varepsilon}$ is almost surely finite. Thus $X_t^{x,\varepsilon}$ is a Brownian motion starting at x, with drift $2\varepsilon b$, and whose fluctuations are scaled by $\varepsilon\sqrt{2A_b}$. According to the Feynman-Kac formula, see Theorem 4.A.11 in section 4.A, the following expectation is the unique solution to the PDE (4.19)

$$u_\varepsilon(x) = \mathbb{E}\left[\exp\left(-\tau^{x,\varepsilon} - \frac{g(X^{x,\varepsilon}_{\tau^{x,\varepsilon}})}{\varepsilon}\right)\right]. \quad (4.23)$$

In particular, u_ε is positive. In the framework of the stochastic approach, Theorem 4.2.12 expresses the convergence of the following soft-minimum

$$\mathbf{u}_\varepsilon(x) = -\varepsilon \ln\left(\mathbb{E}\left[\exp\left(-\tau^{x,\varepsilon} - \frac{g(X^{x,\varepsilon}_{\tau^{x,\varepsilon}})}{\varepsilon}\right)\right]\right), \quad (4.24)$$

towards the minimum (4.14) defining the value function $\mathbf{u}$.

Remark 4.2.13 (Divergence form Laplacian). One may replace in (4.19) the non-divergence form anisotropic Laplacian with the divergence form variant

$$\operatorname{div}(A_b\nabla u) = \operatorname{Tr}(A_b\nabla^2 u) + \langle \operatorname{div}(A_b), \nabla u\rangle,$$

where $\operatorname{div}(A_b)$ denotes column-wise divergence, assuming that A_b is continuously differentiable. Indeed, this amounts to replacing in (4.19) the vector field b defining the first order term with $b_\varepsilon := b - \frac{\varepsilon}{2}\operatorname{div}(A_b)$. This small perturbation is easily handled in the setting of viscosity solutions, and the same limit (4.4) is obtained as $\varepsilon \to 0$.

The divergence form Laplacian is often preferred in applications [CWW13] since it is simpler to implement numerically on some geometries, such as triangulated surfaces using finite elements. Finite element methods may however lack the discrete comparison principle Lemma 4.3.4 used to establish the convergence of the numerical scheme in this paper.

4.3 The numerical scheme

We present a numerical implementation of the linear second order PDE (4.3) based on discrete degenerate elliptic finite differences, on a Cartesian discretization grid. This approach is chosen for the simplicity of its implementation and of the convergence analysis. Alternative discretizations may also be considered, for instance using finite elements on triangulated manifolds, see [CWW13] and Remark 4.2.13.

Throughout this section, we denote by $h > 0$ the grid scale of the Cartesian discretization grid, which is fixed unless otherwise specified, and we define the discrete domain as

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \qquad \overline{\Omega}_h := \Omega_h \cup \partial\Omega. \tag{4.25}$$

In our application, the values of u on $\partial\Omega$ are given by the Dirichlet boundary conditions, and the numerical implementation does not treat them as unknowns. For any $u : \Omega_h \to \mathbb{R}$, any $x \in \Omega_h$ and any $e \in \mathbb{Z}^d$, we define the first order and second order centered finite differences operators as follows: assuming $[x - he,\, x + he] \subset \Omega$

$$\overline{\delta}^e_h u(x) := \frac{u(x + he) - u(x - he)}{2h}, \qquad \Delta^e_h u(x) := \frac{u(x + he) - 2u(x) + u(x - he)}{h^2}. \tag{4.26}$$

If x is adjacent to $\partial\Omega$, then (4.26) may involve values outside the domain Ω_h, and thus be ill-defined. In order to address this issue, we consider $u : \overline{\Omega}_h \to \mathbb{R}$ which is also defined on the domain boundary. The following finite difference expressions make sense for arbitrary $x \in \Omega_h$, $e \in \mathbb{Z}^d$, and they reduce to (4.26) if $[x - he,\, x + he] \subset \Omega$:

$$\overline{\delta}^e_h u(x) := \frac{1}{2}\Big(\frac{u(x + h^e_x e) - u(x)}{h^e_x} - \frac{u(x - h^{-e}_x e) - u(x)}{h^{-e}_x}\Big), \tag{4.27}$$

$$\Delta^e_h u(x) := \frac{2}{h^e_x + h^{-e}_x}\Big(\frac{u(x + h^e_x e) - u(x)}{h^e_x} + \frac{u(x - h^{-e}_x e) - u(x)}{h^{-e}_x}\Big), \tag{4.28}$$

where we denoted

$$h^e_x := \min\{\eta > 0;\, x + \eta e \in \overline{\Omega}_h\}. \tag{4.29}$$

Note that $h^e_x \in]0, h]$ by construction. If $u \in C^4(\overline{\Omega})$ then one has the consistency relation

$$\overline{\delta}^e_h u(x) = \langle \nabla u(x), e\rangle + \mathcal{O}(h^r), \qquad \Delta^e_h u(x) = \langle e, \nabla^2 u(x) e\rangle + \mathcal{O}(h^r),$$

where $r = 2$ if $[x - he, x + he] \subset \overline{\Omega}$, and $r = 1$ otherwise. In the next proposition we obtain, by linear combination, consistent finite differences approximations of linear PDE operators of order one and two.

Proposition 4.3.1. *Let $D \in S_d$, and let $\omega \in \mathbb{R}^d$. Consider weights ρ_i and offsets $e_i \in \mathbb{Z}^d$, for all $1 \le i \le I$, such that*

$$D = \sum_{1 \le i \le I} \rho_i e_i e_i^\top. \tag{4.30}$$

Then for $u \in C^4(\overline{\Omega})$ and $x \in \Omega_h$ one has

$$\sum_{1 \le i \le I} \rho_i\, \overline{\delta}^{e_i}_h u(x)\, e_i = D\nabla u(x) + \mathcal{O}(h^r), \qquad \sum_{1 \le i \le I} \rho_i \Delta^{e_i}_h u(x) = \mathrm{Tr}(D\nabla^2 u(x)) + \mathcal{O}(h^r), \tag{4.31}$$

where $r = 2$ if $[x - he_i, x + he_i] \subset \overline{\Omega}$ for all $1 \le i \le I$, and $r = 1$ otherwise.

As an immediate application, we define a finite difference discretization $\mathcal{L}_h^\varepsilon$ of the linear operator $\mathcal{L}^\varepsilon$ defined in (4.21). For any $u : \overline{\Omega}_h \to \mathbb{R}$ we let

$$\mathcal{L}_h^\varepsilon u = u + 2\varepsilon \sum_{1\le i\le I} \rho_i \langle A_b^{-1} b, e_i\rangle \; \overline{\delta}_h^{e_i} u - \varepsilon^2 \sum_{1\le i\le I} \rho_i \Delta_h^{e_i} u, \tag{4.32}$$

with boundary condition $u = \exp(-g/\varepsilon)$ on $\partial\Omega$. The weights $\rho_i = \rho_i(x)$ and offsets $e_i = e_i(x)$, $1 \le i \le I$, provide a decomposition of the matrix $A_b = A_b(x)$ in the sense of (4.30). Note that for the schemes (4.31) to be well-defined, it is crucial that the offsets involved in (4.30) have integer coordinates, and therefore the similar looking eigenvalue-eigenvector decomposition typically cannot be used since it involves arbitrary unit vectors. Obtaining a suitable decomposition is thus non-trivial in general, and it is also not unique. We rely in this paper on Selling's decomposition, which is defined in dimension $d \in \{2, 3\}$, and has the additional benefit of producing non-negative weights $(\rho_i)_{1\le i\le I}$ and thus a discrete degenerate elliptic scheme, see section 4.3.1 below.

Remark 4.3.2 (Approximation of the gradient, improved reconstruction, following [CWW13]). An approximate gradient $V_h^\varepsilon : \Omega_h \to \mathbb{R}^d$ of the solution u_h^ε of (4.32) can be estimated using (4.31, left):

$$V_h^\varepsilon(x) := A_b(x)^{-1} \sum_{1\le i\le I} \rho_i \, \overline{\delta}_h^{e_i} u_h^\varepsilon(x) e_i, \qquad \mathbf{V}_h^\varepsilon(x) := \frac{-V_h^\varepsilon(x)}{|V_h^\varepsilon(x)|_{A(x)} - \langle b(x), V_h^\varepsilon(x)\rangle}, \tag{4.33}$$

The vector field $\mathbf{V}_h^\varepsilon$ is meant to approximate the gradient of Randers distance $\mathbf{u}$ from the boundary (4.4): it is negatively proportional to V_h^ε, reflecting the fact that logarithmic transformation is decreasing, and is normalized consistently with Randers eikonal equation (4.15). An empirical observation of [CWW13], in the context of isotropic and Riemannian metrics which are special cases of Randers metrics (and using a different discretization), is that $\mathbf{V}_h^\varepsilon$ is for suitable parameters h, ε an excellent approximation of $\nabla\mathbf{u}$. In particular, it can be used for geodesic backtracking via (4.17) and (4.18). Following [CWW13] we may also obtain an empirically improved reconstruction $\mathbf{v}_h^\varepsilon : \Omega_h \to \mathbb{R}$ of the Randers distance by minimizing

$$\sum_{x\in\Omega_h} \sum_{1\le|i|\le I} \rho_i |\delta_h^{e_i} \mathbf{v}(x) - \langle e_i, \mathbf{V}_h^\varepsilon(x)\rangle|^2, \tag{4.34}$$

which is consistent with $\int_\Omega |\nabla\mathbf{v} - \mathbf{V}_h^\varepsilon|_{A_b}^2$, where $\rho_{-i} := \rho_i$ and $e_{-i} := e_i$ for all $1 \le i \le I$, and where the first order upwind finite difference δ_h^e is defined in (4.35). Equations (4.33, left) and (4.34) also make sense if one replaces the weights and offsets $(\rho_i, e_i)_{i=1}^I$ and matrix A_b used in the numerical scheme (4.32), with unit weights and the canonical basis and the identity matrix. However, the latter (and simpler) choice yields slightly less accurate results empirically as evidenced in our numerical experiments section 4.6. In Figures 4.3 and 4.4 we refer to these post-processed distance maps as $\mathbf{u}_h^{A_b}$ and $\mathbf{u}_h^{I_2}$ respectively.

4.3.1 Discrete degenerate ellipticity

Discrete degenerate ellipticity is a counterpart to the degenerate ellipticity property of Hamilton-Jacobi-Bellman PDE operators [CIL92; Obe06], which is at the foundation of the theory of viscosity solutions, see Definition 4.A.1.

Definition 4.3.3 (Discrete degenerate ellipticity [Obe06])**.** Let X be a finite set, and let $\mathbb{U} := \mathbb{R}^X$. A (finite difference) scheme on X is a function $F : \mathbb{U} \to \mathbb{U}$. Such a function can be written in the form

$$Fu(x) := \tilde{F}(x, u(x), (u(x) - u(y))_{y\in X\setminus\{x\}}),$$

and the scheme is said discrete degenerate elliptic (DDE) if $\tilde{F}$ is non-decreasing w.r.t. the second variable, and w.r.t. the third variable componentwise. The scheme is said elliptic if $u \mapsto Fu - \lambda u$ is degenerate elliptic for some $\lambda > 0$.

Similarly to its continuous counterpart, discrete ellipticity implies a comparison principle, used in the proof of the existence and uniqueness of solutions to discretized PDEs, and of their convergence to the continuous solutions as the grid scale is refined section 4.3.3. For completeness, we present the proof of two basic but fundamental properties of discrete elliptic operators, see e.g. [Obe06] for additional discussion.

Lemma 4.3.4 (Discrete comparison principle)**.** *Let F be an elliptic finite differences scheme on a finite set X, and let $u, v : X \to \mathbb{R}$. If $Fu \leq Fv$ on X, then $u \leq v$ on X.*

Proof. Let $x_* \in X$ be such that $u(x_*) - v(x_*)$ is maximal, so that $u(x_*) - u(y) \geq v(x_*) - v(y)$ for all $y \in X$. Assume for contradiction that $u(x_*) > v(x_*)$, otherwise the result is proved. Then, by discrete degenerate ellipticity of $F - \lambda \operatorname{Id}$, we obtain $Fu(x_*) - \lambda u(x_*) \geq Fv(x_*) - \lambda v(x_*)$, thus $0 < \lambda(u(x_*) - v(x_*)) \leq Fu(x_*) - Fv(x_*) \leq 0$, which proves the result by contradiction. □

We say that u is a sub-solution (resp. super-solution, resp. solution) of the scheme F, if $Fu \leq 0$ (resp. $Fu \geq 0$, resp. $Fu = 0$) on X.

Corollary 4.3.5 (Solution to elliptic linear operators)**.** *If F is an affine (i.e. linear plus constant) and elliptic scheme on a finite set X, then there exists a unique solution $u : X \to \mathbb{R}$ to $Fu = 0$.*

Proof. If $Fu = Fv$ on X then $u = v$, by Lemma 4.3.4. Thus $F : \mathbb{R}^X \to \mathbb{R}^X$ is injective, hence by linearity it is bijective, and there exists a unique solution to $Fu = 0$. □

The finite difference schemes considered in this paper (4.27), (4.28), and (4.32) formally involve a function defined on the uncountable set $\overline{\Omega}_h = \Omega_h \cup \partial\Omega$, which does not comply with the finiteness assumption in Definition 4.3.3. This obstruction is only superficial, since only finitely many boundary values of u are actually involved these schemes, for any given $h > 0$. Alternatively, one may consider the Dirichlet boundary values of u as given constants rather than unknown variables in the scheme.

The simplest DDE operator is the opposite $-\delta_h^e$ of the upwind finite difference δ_h^e on Ω_h, where $h > 0$ and $e \in \mathbb{Z}^d$, which is defined as

$$\delta_h^e u(x) := \frac{u(x + he) - u(x)}{h}. \tag{4.35}$$

The operator δ_h^e is modified similarly to (4.27) and (4.28) if $[x, x + he] \not\subset \overline{\Omega}$, and is first order consistent with a directional derivative: for any $u : \overline{\Omega}_h \to \mathbb{R}$ and any $x \in \Omega_h$

$$\delta_h^e u(x) := \frac{u(x + h_x^e e) - u(x)}{h_x^e}, \qquad \delta_h^e u(x) = \langle e, \nabla u(x) \rangle + \mathcal{O}(h). \tag{4.36}$$

The opposite $-\Delta_h^e$ of the second order finite difference operator Δ_h^e is also DDE. The centered finite difference operator $\overline{\delta}_h^e$ is not DDE, but linear combinations with Δ_h^e whose coefficients have suitable signs and obey suitable bounds satisfy this property, as shown in the next lemma. For that purpose, we observe the relations

$$\Delta_h^e u(x) = \frac{2}{h_x^e + h_x^{-e}} \left(\delta_h^e u(x) + \delta_h^{-e} u(x) \right), \qquad \overline{\delta}_h^e u(x) = \frac{1}{2} \left(\delta_h^e u(x) - \delta_h^{-e} u(x) \right). \tag{4.37}$$

Lemma 4.3.6. *Let $e \in \mathbb{Z}^d$, and $h > 0$. The finite difference scheme $-\Delta_h^e$ is unconditionally DDE, and the linear combination $\mu\overline{\delta}_h^e - \lambda\Delta_h^e$ is DDE when $h|\mu| \leq 2\lambda$.*

Proof. In view of (4.37) one has the equality of schemes $\mu\overline{\delta}_h^e - \lambda\Delta_h^e = -\alpha\delta_h^e - \beta\delta_h^{-e}$, where $\alpha : X \to \mathbb{R}$ is defined by $\alpha(x) := 2\lambda/(h_x^e + h_x^{-e}) - \mu/2$ which is non-negative if $h|\mu| \leq 2\lambda$, since $0 < h_x^{\pm e} \leq h$. Likewise $\beta(x) := 2\lambda/(h_x^e + h_x^{-e}) + \mu/2 \geq 0$ if $h|\mu| \leq 2\lambda$. We conclude by observing that DDE schemes form a cone: linear combinations with non-negative coefficients of DDE schemes are DDE. □

Corollary 4.3.7. *The finite difference scheme $\mathcal{L}_h^\varepsilon$ defined by (4.32) is elliptic, with $\lambda = 1$, if $\rho_i \geq 0$ and $h|\langle A_b^{-1}b, e_i\rangle| \leq \varepsilon$ for all $1 \leq i \leq I$.*

Proof. Under these assumptions, the finite difference scheme $u \mapsto \mathcal{L}_h^\varepsilon u - u$ is the sum of the finite difference operators $\varepsilon\rho_i(2\mu_i\overline{\delta}_h^{e_i}u - \varepsilon\Delta_h^{e_i})$ where $\mu_i = \langle A_b^{-1}b, e_i\rangle$, for all $1 \leq i \leq I$. By Lemma 4.3.6, which applies regardless of the fact that ρ_i and e_i depend on the point $x \in \Omega_h$, each of these elementary operators is DDE when $\rho_i \geq 0$ and $h|\mu_i| \leq \varepsilon$. Hence $\mathcal{L}_h^\varepsilon - \mathrm{Id}$ is DDE, and therefore $\mathcal{L}_h^\varepsilon$ is elliptic with $\lambda = 1$ by Definition 4.3.3. □

As announced in the introduction of this section, and in order to benefit from Lemma 4.3.4 and Corollary 4.3.5, we do want the discrete operator $\mathcal{L}_h^\varepsilon$ to be DDE. For that purpose, we introduce Selling's decomposition [Sel74; CS92] of a positive definite matrix $D \in S_d^{++}$, where $d \in \{2, 3\}$, which is efficiently computable numerically via Selling's algorithm. In view of their key role in our numerical scheme, Selling's constructions and some of their properties are presented in more detail in section 4.B.

Theorem 4.3.8 (Selling [Sel74], this version [Mir18])**.** *Let $D \in S_d^{++}$, where $d \in \{2, 3\}$. Then there exists non-negative weights $\rho_i \geq 0$, and offsets $e_i \in \mathbb{Z}^d$, where $1 \leq i \leq I := d(d+1)/2$, such that*

$$D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^\top, \qquad\qquad |e_i| \leq 2C_d\,\mu(D),\ \forall 1 \leq i \leq I,$$

where $C_2 = 2$, $C_3 = 2\sqrt{3}$, and $\mu(D) := \sqrt{\|D\|\|D^{-1}\|}$ is the anisotropy ratio of D.

In the rest of this section, we assume that the weights and offsets $(\rho_i(x), e_i(x))_{i=1}^I$ used to define the scheme $\mathcal{L}_h^\varepsilon$, see (4.32), are obtained from Selling's decomposition of the matrix $A_b(x)$, for all $x \in \Omega_h$. For the sake of readability, the dependency of ρ_i and e_i w.r.t. the base point x is often left implicit in the equations. The following proposition, stated without proof, immediately follows from Corollary 4.3.7 and Theorem 4.3.8.

Proposition 4.3.9. *The scheme $\mathcal{L}_h^\varepsilon$ is elliptic provided that $Ch \leq \varepsilon$, where*

$$C := 2C_d \max_{x \in \overline{\Omega}} \mu(A_b(x))\,|A_b^{-1}(x)b(x)|.$$

The construction of finite difference schemes for linear and semi-linear PDEs using Selling's algorithm, and the compatibility conditions ensuring the DDE property, are discussed in more detail in Chapter 2. Finally, let us mention an alternative discretization of the PDE operator $\mathcal{L}^\varepsilon$ defined in (4.21), using upwind finite differences for the first order term, which is unconditionally stable but has a lower consistency order

$$\mathcal{L}_h^{\varepsilon,+}u = u - 2\varepsilon \sum_{1 \leq j \leq d} |\langle b, f_j\rangle|\delta_h^{-\sigma_j f_j}u - \varepsilon^2 \sum_{1 \leq i \leq I} \rho_i \Delta_h^{e_i}u, \tag{4.38}$$

where $(f_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$, and σ_j is the sign of $\langle b, f_j\rangle$.

4.3.2 Logarithmic transformation

We use a logarithmic transformation of the unknown to study the convergence of the solutions to the discrete schemes (4.32) and (4.38) as the relaxation parameter ε and the grid scale h tend to zero suitably, mimicking the approach used in the continuous case, see section 4.2.4. Our first step is to describe the effect of the logarithmic/exponential transformation on a finite difference scheme.

Proposition 4.3.10. *Let $h > 0$ and $\varepsilon > 0$. Let F be a DDE scheme on $\overline{\Omega}_h$, such that $Fu(x)$ is a* linear *function of u for all $x \in \Omega_h$, with boundary condition $u = \exp(-g/\varepsilon)$ on $\partial\Omega$, where $u : \overline{\Omega}_h \to \mathbb{R}$. We define the exponentially transformed scheme $\mathbf{F}^\varepsilon$ as follows:*

$$\begin{aligned}\mathbf{F}^\varepsilon\mathbf{u}(x) &:= -e^{\frac{\mathbf{u}(x)}{\varepsilon}}\,\big[Fe^{\frac{-\mathbf{u}}{\varepsilon}}\big](x) \\ &= \tilde{F}\Big(x, -1, \Big[\exp\big(\frac{\mathbf{u}(x)-\mathbf{u}(y)}{\varepsilon}\big) - 1\Big]_{y\in X\setminus\{x\}}\Big),\end{aligned} \tag{4.39}$$

for any $x \in \Omega_h$, with boundary condition $\mathbf{u} = g$ on $\partial\Omega$, where $\mathbf{u} : \overline{\Omega}_h \to \mathbb{R}$. The scheme $\mathbf{F}^\varepsilon$ is DDE, and furthermore if $\mathbf{u}$ is a sub-solution (resp. super-solution) of $\mathbf{F}^\varepsilon$, then $u := \exp(-\mathbf{u}/\varepsilon)$ is a super-solution (resp. sub-solution) of F.

Proof. The two expressions of $\mathbf{F}^\varepsilon\mathbf{u}(x)$ given in (4.39), where $x \in \Omega_h$, are equivalent in view of the linearity of $\tilde{F}$. The discrete degenerate ellipticity of $\mathbf{F}^\varepsilon$ follows from the same property of F, and from the fact that $t \in \mathbb{R} \mapsto \exp(t/\varepsilon) - 1$ is non-decreasing. □

We use the scheme unknown transformation $u = \exp(-\mathbf{u}/\varepsilon)$, which is classical in the study of relations between the heat, Poisson, and eikonal equations [Var67; CWW13]. However, since the mapping $t \mapsto \exp(-t/\varepsilon)$ is decreasing, it exchanges the notions of sub-solutions and super-solutions, see Proposition 4.3.10. The exponentially transformed upwind finite difference is denoted $\boldsymbol{\delta}_h^{e,\varepsilon}$, and reads

$$\boldsymbol{\delta}_h^{e,\varepsilon}\mathbf{u}(x) = \frac{1}{h}\Big(1 - \exp\big(\frac{\mathbf{u}(x)-\mathbf{u}(x+he)}{\varepsilon}\big)\Big), \tag{4.40}$$

where $x \in \Omega_h$, $e \in \mathbb{Z}^d$, and assuming $[x, x+he] \subset \overline{\Omega}$. Otherwise replace h with h_x^e in the above expression, see (4.29). The next lemma approximates (4.40) in terms of the derivatives of $\mathbf{u}$.

Lemma 4.3.11. *Let $\mathbf{u} \in C^3(\overline{\Omega})$ and $0 < h \le \varepsilon \le 1$. Then for any $x \in \Omega_h$, and bounded $e \in \mathbb{Z}^d$,*

$$\boldsymbol{\delta}_h^{e,\varepsilon}\mathbf{u}(x) = \frac{1}{\varepsilon}\langle\nabla\mathbf{u}(x), e\rangle + \frac{h}{2\varepsilon}\langle e, \nabla^2\mathbf{u}(x)e\rangle - \frac{h}{2\varepsilon^2}\langle\nabla\mathbf{u}(x), e\rangle^2 + \frac{h^2}{6\varepsilon^3}\langle\nabla\mathbf{u}(x), e\rangle^3 + \mathcal{O}\big(\frac{h^2}{\varepsilon^2} + \frac{h^3}{\varepsilon^4}\big), \tag{4.41}$$

assuming $[x, x+he] \subset \overline{\Omega}$. Otherwise, replace h with h_x^ε in the above expression.

Proof. The announced result immediately follows from (4.40) and the Taylor expansion $1 - \exp(-s) = s - \frac{1}{2}s^2 + \frac{1}{6}s^3 + \mathcal{O}(s^4)$, where s is defined by $\varepsilon s = \mathbf{u}(x+he) - \mathbf{u}(x) = h\langle\nabla\mathbf{u}(x), e\rangle + \frac{1}{2}h^2\langle e, \nabla^2\mathbf{u}(x)e\rangle + \mathcal{O}(h^3)$. □

The exponentially transformed second order and first order centered finite difference operators are denoted $\overline{\Delta}_h^{e,\varepsilon}$ and $\overline{\boldsymbol{\delta}}_h^{e,\varepsilon}$, and their Taylor expansion is deduced from that of $\boldsymbol{\delta}_h^{e,\varepsilon}$. The assumption $0 < h \le \varepsilon \le 1$ of Lemma 4.3.11 serves to eliminate spurious negligible terms in the Taylor expansion, and is asymptotically satisfied in convergence analysis Theorem 4.3.18 which requires $\varepsilon \to 0$ and $h/\varepsilon \to 0$. Note that if $\varepsilon = \mathcal{O}(\sqrt{h})$, as considered in Corollary 4.3.14 below, then the remainder in (4.41) (resp. (4.42) and (4.44) below) simplifies to $\mathcal{O}(h^3/\varepsilon^4)$ (resp. $\mathcal{O}(h^r/\varepsilon^{2+r})$ and $\mathcal{O}(h^r/\varepsilon^r)$).

Corollary 4.3.12. *Under the assumptions of Lemma 4.3.11, one has*

$$\overline{\Delta}_h^{e,\varepsilon}\mathbf{u}(x) = \frac{1}{\varepsilon}\langle e, \nabla\mathbf{u}(x)e\rangle - \frac{1}{\varepsilon^2}\langle\nabla\mathbf{u}(x), e\rangle^2 + \mathcal{O}\big(\frac{h}{\varepsilon^2} + \frac{h^r}{\varepsilon^{2+r}}\big), \tag{4.42}$$
$$\overline{\boldsymbol{\delta}}_h^{e,\varepsilon}\mathbf{u}(x) = \frac{1}{\varepsilon}\langle\nabla\mathbf{u}(x), e\rangle + \mathcal{O}\big(\frac{h^r}{\varepsilon^{1+r}}\big),$$

where $r = 2$ if $[x - he, x + he] \subset \overline{\Omega}$, and $r = 1$ otherwise.

Proof. The operators $\overline{\Delta}_h^{e,\varepsilon}$ and $\overline{\boldsymbol{\delta}}_h^{e,\varepsilon}$ can be expressed in terms of the corresponding upwind finite difference operators $\boldsymbol{\delta}_h^{\pm e,\varepsilon}$, similarly to their original counterparts (4.37). The announced result follows by inserting the Taylor expansion obtained in Lemma 4.3.11. In the case where $[x - he, x + he] \subset \overline{\Omega}$, the expansion of $\overline{\Delta}_h^{e,\varepsilon} = \frac{1}{h}(\boldsymbol{\delta}_h^{e,\varepsilon} + \boldsymbol{\delta}_h^{-e,\varepsilon})$ benefits form the cancellation of the term $\langle\nabla\mathbf{u}(x), e\rangle^3$ in (4.41) which is anti-symmetric w.r.t. e, and likewise the expansion of $\overline{\boldsymbol{\delta}}_h^{e,\varepsilon} = \frac{1}{2}(\boldsymbol{\delta}_h^{e,\varepsilon} - \boldsymbol{\delta}_h^{-e,\varepsilon})$ benefits from the cancellation of the terms $\langle\nabla\mathbf{u}, e\rangle^2$ and $\langle e, \nabla^2\mathbf{u}\, e\rangle$ in (4.41) which are symmetric w.r.t. e. □

Consistently with the continuous case (4.21), we denote by $\mathcal{S}_h^\varepsilon$ the exponential transformation of the finite differences scheme $\mathcal{L}_h^\varepsilon$ defined by (4.32). In other words, following Proposition 4.3.10

$$\mathcal{S}_h^\varepsilon\mathbf{u} := -e^{\frac{\mathbf{u}}{\varepsilon}}\mathcal{L}_h^\varepsilon e^{-\frac{\mathbf{u}}{\varepsilon}} \tag{4.43}$$

on Ω_h, with boundary condition $\mathbf{u} = g$ on $\partial\Omega$.

Proposition 4.3.13 (Consistency with the regularized eikonal equation)**.** *For any $\mathbf{u} \in C^3(\overline{\Omega})$, any $0 < h \le \varepsilon \le 1$, and any $x \in \Omega_h$ one has*

$$\mathcal{S}_h^\varepsilon\mathbf{u}(x) = \mathcal{S}^\varepsilon\mathbf{u}(x) + \mathcal{O}(h + h^r/\varepsilon^r), \quad \textit{where} \quad \mathcal{S}^\varepsilon\mathbf{u} := |\nabla\mathbf{u}|^2_{A_b} + 2\langle b, \nabla\mathbf{u}\rangle - 1 - \varepsilon\,\mathrm{Tr}(A_b\nabla^2\mathbf{u}), \tag{4.44}$$

and where $r = 2$ if $[x - he_i, x + he_i] \subset \overline{\Omega}$ for all $1 \le i \le I$, and $r = 1$ otherwise.

Proof. Denoting $\mu_i := \rho_i\langle A_b^{-1}b, e_i\rangle$ we obtain as announced,

$$\begin{aligned}
\mathcal{S}_h^\varepsilon\mathbf{u}(x) &= 1 + 2\varepsilon\sum_{1\le i\le I}\mu_i\overline{\boldsymbol{\delta}}_h^{e_i}\mathbf{u}(x) - \varepsilon^2\sum_{1\le i\le I}\rho_i\overline{\Delta}_h^{e_i}\mathbf{u}(x)\\
&\approx 1 + 2\sum_{1\le i\le I}\mu_i\langle e_i, \nabla\mathbf{u}(x)\rangle + \sum_{1\le i\le I}\rho_i\langle e_i, \nabla\mathbf{u}(x)\rangle^2 - \varepsilon\sum_{1\le i\le I}\rho_i\langle e_i\nabla^2\mathbf{u}(x)e_i\rangle\\
&= 1 + 2\Big\langle\sum_{1\le i\le I}\mu_i e_i,\, \nabla\mathbf{u}(x)\Big\rangle + \mathrm{Tr}\Big(\big(\nabla\mathbf{u}(x)\nabla\mathbf{u}(x)^\top - \varepsilon\nabla^2\mathbf{u}(x)\big)\sum_{1\le i\le I}\rho_i e_i e_i^\top\Big)\\
&= 1 + 2\langle b, \nabla\mathbf{u}(x)\rangle + |\nabla\mathbf{u}(x)|^2_{A_b(x)} - \varepsilon\,\mathrm{Tr}(A_b(x)\nabla^2\mathbf{u}(x)),
\end{aligned}$$

where $\approx$ denotes equality up to a $\mathcal{O}(h + h^r/\varepsilon^r)$ error. □

We obtain a consistency order of $2/3$ in the domain interior, and $1/2$ close to the boundary, by choosing ε as an optimal power of h (respectively $\varepsilon = h^{2/3}$ and $\varepsilon = h^{1/2}$).

Corollary 4.3.14 (Consistency with the eikonal equation)**.** *For any $u \in C^3(\overline{\Omega})$, any $0 < h \le \varepsilon \le 1$, and any $x \in \Omega_h$ one has*

$$\mathcal{S}_h^{h^\alpha}\mathbf{u}(x) = \mathcal{S}\mathbf{u}(x) + \mathcal{O}(h^\alpha), \qquad \textit{where} \quad \mathcal{S}\mathbf{u} := |\nabla\mathbf{u}|^2_{A_b} + 2\langle b, \nabla\mathbf{u}\rangle - 1,$$

and where $\alpha = 2/3$ if $[x - he_i, x + he_i] \subset \overline{\Omega}$ for all $1 \le i \le I$, and $\alpha = 1/2$ otherwise.

Proof. One has $\mathcal{S}^\varepsilon \mathbf{u} = \mathcal{S}\mathbf{u} - \varepsilon \operatorname{Tr}(A_b \nabla^2 u)$, and therefore $\mathcal{S}_h^\varepsilon \mathbf{u}(x) = \mathcal{S}\mathbf{u} + \mathcal{O}(\varepsilon + h + h^r \varepsilon^{-r})$, where r is defined pointwise as in Proposition 4.3.13. Observing that $\alpha = r/(1+r)$, and inserting $\varepsilon = h^\alpha$ in this expression, one obtains the announced result. □

While the above suggests choosing a different value of ε at points that are close to the boundary, in our numerical experiments we use for simplicity a single value of ε on the whole domain. The theoretical analysis of convergence rates of the method, and of the actual effect of a differentiated choice of ε on those rates, is not developped in this paper and is an opportunity for future work.

The upwind scheme $\mathcal{L}_h^{\varepsilon,+}$ obeys Proposition 4.3.13 but with $r = 1$ over all Ω_h, and likewise Corollary 4.3.14 but with $\alpha = 1/2$ over all Ω_h.

Note that the choice $\varepsilon = h^\alpha$ with $\alpha = \frac{r}{1+r}$, considered in Corollary 4.3.14, minimizes the error term $\sigma(h,\varepsilon) := \varepsilon + h + h^r \varepsilon^{-r}$ up to a fixed multiplicative constant. Indeed $\sigma(h, h^\alpha) = \mathcal{O}(h^\alpha)$ whereas $\sigma(h,\varepsilon) = \varepsilon + h + h^r \varepsilon^{-r} \geq \alpha\varepsilon + (1-\alpha) h^r \varepsilon^{-r} \geq \varepsilon^\alpha (h^r/\varepsilon^r)^{1-\alpha} = h^\alpha$, where the concavity of the logarithm was used for the second inequality. The parameter scaling $h = c\varepsilon$, where $c > 0$ is a small but fixed positive constant, is commonly considered in applications [CWW13] and appears to produce usable results in practice, but is not consistent asymptotically since $\sigma(h, ch) \to c^r$. In the simplified setting where $d = 1$, $A = 1$ and $b = 0$, one easily checks that $\mathcal{S}_h^\varepsilon$ admits the solution $\mathbf{u}(x) = \lambda x$ (with suitable boundary conditions) where the slope λ obeys

$$e^{c\lambda} + e^{-c\lambda} = 2 + c^2 \qquad \text{thus } |\lambda| = 1 - c^2/24 + \mathcal{O}(c^4), \tag{4.45}$$

where $c = h/\varepsilon$. The correct slope $|\lambda| = 1$ is thus only obtained as $c = h/\varepsilon \to 0$.

4.3.3 Convergence

We establish the convergence of the logarithmically transformed solution to the numerical scheme $\mathcal{L}_h^\varepsilon$, towards the solution of Randers eikonal equation as $\varepsilon \to 0$ and $h/\varepsilon \to 0$, see Theorem 4.3.18 which was announced in the introduction. The proof follows the lines of [BS91, Theorem 2.1], and requires some preliminary steps establishing the stability and consistency of the proposed scheme. The arguments apply without modification to the less accurate but unconditionally stable $\mathcal{L}_h^{\varepsilon,+}$. Note that, formally, the schemes $\mathcal{S}_h^\varepsilon$ and $\mathcal{L}_h^\varepsilon$ are defined over $\overline{\Omega}_h := \Omega_h \cup \partial\Omega$. In particular $\mathcal{S}_h^\varepsilon \mathbf{u}(x) = \mathbf{u}(x) - g(x)$ and $\mathcal{L}_h^\varepsilon u(x) = u(x) - \exp(-g(x)/\varepsilon)$ for all $x \in \partial\Omega$ and $u, \mathbf{u} : \overline{\Omega}_h \to \mathbb{R}$.

Lemma 4.3.15. *The scheme $\mathcal{S}_h^\varepsilon$ admits a constant sub-solution $\overline{\mathbf{u}} : \overline{\Omega}_h \to \mathbb{R}$ defined as*

$$\overline{\mathbf{u}}(x) := g_{\min}, \qquad \text{where} \quad g_{\min} := \min_{y \in \partial\Omega} g(y),$$

and for any $p \in \mathbb{R}^d$ with $|p|$ sufficiently large and $(\varepsilon, h/\varepsilon)$ small enough, a super-solution $\underline{\mathbf{u}} : \overline{\Omega}_h \to \mathbb{R}$ defined as the affine map

$$\underline{\mathbf{u}}(x) := \langle p, x \rangle + c_{\max}, \qquad \text{where} \quad c_{\max} := \max_{y \in \partial\Omega} \big(g(y) - \langle p, y \rangle \big).$$

Proof. Case of the sub-solution. One has $\mathcal{S}_h^\varepsilon \overline{\mathbf{u}}(x) = -1$ for all $x \in \Omega_h$, in view of (4.32) and (4.39). In addition $\mathcal{S}_h^\varepsilon \overline{\mathbf{u}}(x) = g_{\min} - g(x) \leq 0$ for all $x \in \partial\Omega$, hence $\overline{\mathbf{u}}$ is a sub-solution of $\mathcal{S}_h^\varepsilon$.

Case of the super-solution. If $|p|$ is sufficiently large, then for all $x \in \overline{\Omega}$

$$|p|^2_{A_b(x)} + 2\langle b(x), p \rangle - 1 \geq c_0 > 0. \tag{4.46}$$

Indeed, recall that the matrix field $A_b : \overline{\Omega} \to S_d^{++}$ is pointwise positive definite (4.10), and continuous. Then by Proposition 4.3.13, $\mathcal{S}_h^\varepsilon \underline{\mathbf{u}}(x) \geq c_0 + \mathcal{O}(h + h^r/\varepsilon^r)$ for all $x \in \Omega_h$, which is non-negative for $(\varepsilon, h/\varepsilon)$ small enough. In addition $\mathcal{S}_h^\varepsilon \underline{\mathbf{u}}(x) = c_{\max} + \langle p, x \rangle - g(x) \geq 0$ for all $x \in \partial\Omega$, hence $\underline{\mathbf{u}}$ is a sub-solution of $\mathcal{S}_h^\varepsilon$. □

As a consequence, we prove in the next lemma that the scheme $\mathcal{S}_h^\varepsilon$ admit a unique solution, uniformly bounded as $(\varepsilon, h/\varepsilon) \to 0$.

Corollary 4.3.16 (Stability)**.** *For sufficiently small $(\varepsilon, h/\varepsilon)$, the scheme $\mathcal{L}_h^\varepsilon$ admits a unique solution u_h^ε, which is positive, and $\mathcal{S}_h^\varepsilon$ admits a unique solution $\mathbf{u}_h^\varepsilon$, which obeys $\mathbf{u}_h^\varepsilon = -\varepsilon \ln u_h^\varepsilon$ and satisfies $\underline{\mathbf{u}} \le \mathbf{u}_h^\varepsilon \le \overline{\mathbf{u}}$ on $\overline{\Omega}_h$, where $\underline{\mathbf{u}}$ and $\overline{\mathbf{u}}$ are from Lemma 4.3.15.*

Proof. By Proposition 4.3.10, the maps $\underline{u}^\varepsilon := \exp(-\overline{\mathbf{u}}/\varepsilon)$ and $\overline{u}^\varepsilon := \exp(-\underline{\mathbf{u}}/\varepsilon)$, where $\overline{\mathbf{u}}$ and $\underline{\mathbf{u}}$ are from Lemma 4.3.15, are respectively a super-solution and a sub-solution to the scheme $\mathcal{L}_h^\varepsilon$, which is elliptic by Proposition 4.3.9. Since that scheme is also linear, it admits a unique solution u_h^ε by Corollary 4.3.5, obeying $\overline{u}^\varepsilon \le u_h^\varepsilon \le \underline{u}^\varepsilon$ by Lemma 4.3.4. Note that Corollary 4.3.5 and Lemma 4.3.4 apply here regardless of the fact that the domain $\overline{\Omega}_h = \Omega_h \cup \partial\Omega$ is infinite, because the finite difference scheme $\mathcal{L}_h^\varepsilon$ only uses finitely many boundary values. We conclude that u_h^ε is positive since $\underline{u}^\varepsilon$ is positive, that $\mathbf{u}_h^\varepsilon := -\varepsilon \ln u_h^\varepsilon$ is the unique solution to $\mathcal{S}_h^\varepsilon$ by Proposition 4.3.10, and that $\underline{\mathbf{u}} \le \mathbf{u}_h^\varepsilon \le \overline{\mathbf{u}}$ on $\overline{\Omega}_h$ by monotony of the logarithm. The result follows. □

Lemma 4.3.17 (Consistency up to the boundary)**.** *For any $\varphi \in C^3(\overline{\Omega})$ and any $x \in \overline{\Omega}$ one has*

$$\limsup_{\substack{(\varepsilon,h/\varepsilon)\to 0, \xi\to 0 \\ y\in\overline{\Omega}_h,\, y\to x}} \mathcal{S}_h^\varepsilon[\varphi+\xi](y) \le \begin{cases} \mathcal{S}\varphi(x) & \text{if } x \in \Omega, \\ \max\{\mathcal{S}\varphi(x), \varphi(x) - g(x)\} & \text{if } x \in \partial\Omega. \end{cases}$$

$$\liminf_{\substack{(\varepsilon,h/\varepsilon)\to 0, \xi\to 0 \\ y\in\overline{\Omega}_h,\, y\to x}} \mathcal{S}_h^\varepsilon[\varphi+\xi](y) \ge \begin{cases} \mathcal{S}\varphi(x) & \text{if } x \in \Omega, \\ \min\{\mathcal{S}\varphi(x), \varphi(x) - g(x)\} & \text{if } x \in \partial\Omega. \end{cases}$$

Proof. For any $h > 0$, $x \in \Omega_h$, and $\xi \in \mathbb{R}$, one has by Proposition 4.3.13

$$\mathcal{S}_h^\varepsilon[\varphi+\xi](x) = \mathcal{S}_h^\varepsilon\varphi(x) = \mathcal{S}\varphi(x) + \mathcal{O}(\varepsilon + h + (h/\varepsilon)^r),$$

where $r \in \{1, 2\}$. In particular $r \ge 1$ and therefore $\varepsilon + (h/\varepsilon)^r \to 0$ as $h \to 0$. The announced result follows from this observation, and from the uniform continuity of the mapping $x \in \overline{\Omega} \mapsto \mathcal{S}\varphi(x) := |\nabla\varphi(x)|^2_{A_b(x)} + 2\langle b, \nabla\varphi(x)\rangle - 1$. □

Theorem 4.3.18 (Convergence)**.** *As $\varepsilon \to 0$ and $h/\varepsilon \to 0$ the quantity $\mathbf{u}_h^\varepsilon := -\varepsilon \ln u_h^\varepsilon$, where $\mathcal{L}_h^\varepsilon u_h^\varepsilon = 0$, converges uniformly on compact subsets of Ω to the viscosity solution $\mathbf{u}$ of (4.15).*

Proof. Define for all $x \in \overline{\Omega}$

$$\overline{\mathbf{v}}(x) := \limsup_{(\varepsilon,h/\varepsilon)\to 0,\, y\to x} \mathbf{u}_h^\varepsilon(x) \quad \Big(= \sup\big\{ \limsup_{n\to\infty} u_{h_n}^{\varepsilon_n}(y_n);\ (\varepsilon_n, h_n/\varepsilon_n) \to 0,\ y_n \to x,\ y_n \in \overline{\Omega}_{h_n} \big\} \Big),$$

and likewise $\underline{\mathbf{v}}(x) := \liminf \mathbf{u}_h(x)$ as $(\varepsilon, h/\varepsilon) \to 0$ and $y \to x$. By Corollary 4.3.16, $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$ are well-defined and bounded : $\underline{\mathbf{u}} \le \underline{\mathbf{v}} \le \overline{\mathbf{v}} \le \overline{\mathbf{u}}$ on $\overline{\Omega}$ where $\overline{\mathbf{u}}$ and $\underline{\mathbf{u}}$ are from Lemma 4.3.15. By Lemma 4.3.17 and following the proof of [BS91, Theorem 2.1], $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$ are respectively a sub-solution and a super-solution to the operator $\mathcal{S}$, or equivalently to (4.15).

By the continuous comparison principle Theorem 4.A.8, one has $\overline{\mathbf{v}} \le \mathbf{u}_* \le \mathbf{u}^* \le \underline{\mathbf{v}}$ on Ω, where $\mathbf{u}_*(x) := \liminf_{y\to x} \mathbf{u}(y)$ and $\mathbf{u}^*(x) := \limsup_{y\to x} \mathbf{u}(y)$ are the lower and upper semi-continuous envelopes of the solution $\mathbf{u}$ of (4.15). By definition $\overline{\mathbf{v}} \ge \underline{\mathbf{v}}$ on $\overline{\Omega}$, thus $\overline{\mathbf{v}} = \mathbf{u} = \underline{\mathbf{v}}$ on Ω, and the locally uniform convergence follows from the definitions of $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$. □

4.4 Randers distance from a point

In this section, we adapt the numerical scheme presented in section 4.3 so as to compute Randers distance from a point source, instead of the distance to the boundary. Point sources appear to be the most common setting in applications [CWW13; YC16; Yan+18]. However the convergence of the numerical method in this case did not appear to be backed by theory, not least because the corresponding PDE is ill posed, see Remark 4.4.4. To our knowledge, the convergence results of this section Theorems 4.4.1 and 4.4.2 are also new for isotropic and Riemannian metrics, which are special cases of Randers metrics of the form $\mathcal{F}_x(v) = c(x)|v|$ and $\mathcal{F}_x(v) = |v|_{M(x)}$, where $c : \overline{\Omega} \to \mathbb{R}_{++}$ and $M : \overline{\Omega} \to S_d^{++}$, and thus validate previous numerical practice.

We assume that the domain Ω is connected, and contains the origin which w.l.o.g. is the point source of interest, in addition to the previously assumed boundedness and $W^{3,\infty}$ boundary. For all $\varepsilon > 0$, $h > 0$, and $u : \overline{\Omega}_h \to \mathbb{R}$ we let

$$\tilde{\mathcal{L}}_h^\varepsilon u(x) = \begin{cases} \mathcal{L}_h^\varepsilon u(x) & \text{if } x \in \Omega_h \setminus \{0\}, \\ u(x) - 1 & \text{if } x = 0, \\ u(x) & \text{if } x \in \partial\Omega. \end{cases} \tag{4.47}$$

The main result of this section, Theorem 4.4.1 below, justifies the use of the Poisson method, i.e. solving the linear scheme $\tilde{\mathcal{L}}_h^\varepsilon$, to approximate Randers geodesic distance from the origin.

Theorem 4.4.1. *The solution to $\tilde{\mathcal{L}}_h^\varepsilon u_h^\varepsilon = 0$ obeys, locally uniformly in $\Omega \ni x$*

$$-\varepsilon \ln u_h^\varepsilon(x) \to \mathrm{dist}_{\mathcal{F}}(0, x), \qquad \text{as } (\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0.$$

Note that $\tilde{\mathcal{L}}_h^\varepsilon$ is a discrete degenerate elliptic operator when h/ε is sufficiently small, see Proposition 4.3.9, hence it does admit a unique solution by Corollary 4.3.5. Under the same conditions, the matrix of $\mathcal{L}_h^\varepsilon$ is invertible.

Theorem 4.4.2. *Denote by $L_h^\varepsilon \in \mathbb{R}^{\Omega_h \times \Omega_h}$ the matrix of the linear operator $\mathcal{L}_h^\varepsilon$ on Ω_h, with null boundary conditions on $\partial\Omega$. Then locally uniformly on $\Omega \times \Omega \ni (x, y)$ one has*

$$-\varepsilon \ln[(L_h^\varepsilon)_{xy}^{-1}] \to \mathrm{dist}_{\mathcal{F}}(x, y), \qquad \text{as } (\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0.$$

As evidenced by the constraint $\varepsilon \ln h \to 0$, Theorems 4.4.1 and 4.4.2 have no immediate continuous counterparts, see also Remark 4.4.4. Contrast this with the smooth boundary case, where Theorem 4.2.12 corresponds to Theorem 4.3.18 with $h = 0$. The proofs are presented in the rest of this section. In the case of Theorem 4.4.1, it consists in building sub-solutions and a super-solutions to the operator $\tilde{\mathcal{L}}_h^\varepsilon$, on disk or ring domains around the origin depending on the problem scales h, ε and r, where the radius $r > 0$ is fixed but small, see sections 4.4.1 to 4.4.3. Sub-solutions (resp. super-solutions) over these sub-domains are glued together using the following lemma, which immediately follows from the DDE property Definition 4.3.3.

Lemma 4.4.3. *Let F be a DDE scheme on a finite set X, let $x \in X$, and let $u, v : X \to \mathbb{R}$. If $Fu(x) \le 0$ and either ($u(x) \ge v(x)$ or $Fv(x) \le 0$), then $F[\max\{u, v\}](x) \le 0$. Likewise if $Fu(x) \ge 0$ and either ($u(x) \le v(x)$ or $Fv(x) \ge 0$), then $F[\min\{u, v\}](x) \ge 0$.*

Remark 4.4.4 (Continuous setting). The numerical scheme (4.47) does not discretize a well posed PDE. Indeed, Dirichlet boundary conditions cannot be enforced at isolated points of elliptic PDEs in dimension $d \ge 2$. The most closely related well posed PDE is

$$\mathcal{L}^\varepsilon u(x) = \delta_0(x) \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega,$$

where δ_0 denotes the Dirac mass at the origin. This equation admits a solution [Cas86, Theorem 4] in the Sobolev space $W^{1,s}(\Omega)$ where $s < d/(d-1)$, in dimension $d \in \{2,3\}$. The solution is unbounded near 0. We do not further discuss this approach, which belongs to a framework distinct from the setting of viscosity solutions considered in this paper.

Remark 4.4.5 (Heat method). In the Riemannian case ($\omega = 0$) an alternative approach to geodesic distance computation from a point source relies on the short time asymptotics of the heat kernel

$$-4t \ln u(t,x) = \mathrm{dist}_{\mathcal{F}}(x_*, x)^2 + o(1), \qquad \text{where } \partial_t u = \mathrm{div}(D\nabla u), \tag{4.48}$$

and $u(0,\cdot) = \delta_{x_*}$ is the Dirac mass at the source point [Var67]. Numerically, the heat equation is solved over a short time interval, using a series of implicit time steps, each of which is equivalent to a Poisson equation [CWW13]. To the extent of our knowledge, solving a single Poisson equation is the preferred over the heat method in applications, since it is computationally less expensive, and less susceptible to raise floating point underflow errors, in addition to being more general in view of the extension Randers metrics presented in this paper. An advantage of the heat equation is however that it allows efficient implementations of optimal transport with quadratic cost [Sol+15] in the spirit of section 4.5.

A natural generalization of (4.48, right) to Finsler manifolds [OS09] is

$$\partial_t u(t,x) = \mathrm{div}(\partial_v H(x, \nabla u(t,x))), \qquad \text{where } H(x,v) = \tfrac{1}{2} F_x^*(v)^2, \tag{4.49}$$

with again $u(0,\cdot) = \delta_{x_*}$. This PDE can be reformulated as a gradient flow, in two different manners [OS09]. In this setting and under suitable assumptions, the heat kernel asymptotics (4.48, left) extend to Finsler manifolds, see [OS09, Example 5.5]. However, discretizing the non-linear and time dependent PDE (4.49) is non-trivial, and also defeats the purpose of this paper which is to consider *linear* schemes for Randers distance computation. (If non-linear PDEs are considered, then one may as well solve Randers eikonal PDE (4.5) directly, see [Mir14; Mir19].)

Notations The Euclidean ball, its boundary the sphere, and its intersection with the grid, defined for each center $x \in \mathbb{R}^d$, radius $r > 0$ and grid scale $h > 0$, are denoted

$$\mathbb{B}(x,r) := \{y \in \mathbb{R}^d; |y-x| < r\}, \qquad \mathbb{S}(x,r) := \partial\mathbb{B}(x,r), \qquad \mathbb{B}_h(x,r) := \mathbb{B}(x,r) \cap h\mathbb{Z}^d,$$

with the convention $\mathbb{B}(r) := \mathbb{B}(0,r)$, $\mathbb{S}(r) := \mathbb{S}(0,r)$, $\mathbb{B}_h(r) := \mathbb{B}_h(0,r)$. We introduce constants $0 < c_{\mathcal{F}} \le C_{\mathcal{F}}$ and $R_{\mathcal{F}}$, which exist by Lemma 4.2.4, such that for all $x, y \in \overline{\Omega}$

$$c_{\mathcal{F}}|x-y| \le \mathrm{dist}_{\mathcal{F}}(x,y) \le C_{\mathcal{F}}|x-y|, \qquad \mathrm{dist}_{\mathcal{F}}(x,y) \le R_{\mathcal{F}}. \tag{4.50}$$

Recall that the numerical scheme $\mathcal{L}_h^\varepsilon$ is defined in terms of a Lipschitz symmetric matrix field A and vector field b which are the parameters of the dual Randers metric. Selling's decomposition of $A_b := A - bb^\top$, see (4.9), which is uniformly positive definite, is denoted

$$A_b(x) = \sum_{1 \le i \le I} \rho_i(x) e_i e_i^\top, \qquad \text{where } |e_i| \le R_{\mathcal{S}},\ 1 \le i \le I, \tag{4.51}$$

where the bound $R_{\mathcal{S}}$ on the offsets exists in view of Theorem 4.3.8, and I is a suitable integer. The shorthand "$C = C(M_{\mathcal{F}})$" means that a constant C, appearing in an estimate, can be expressed or bounded in terms of the following problem parameters

$$M_{\mathcal{F}} := \max\{c_{\mathcal{F}}^{-1}, C_{\mathcal{F}}, R_{\mathcal{F}}, R_{\mathcal{S}}, \|A\|_\infty, \|b\|_\infty, \|A_b^{-1}\|_\infty, \mathrm{Lip}(A_b)\},$$

where $\|A\|_\infty := \sup\{\|A(x)\|;\ x \in \overline{\Omega}\}$, and $\mathrm{Lip}(A_b)$ is the Lipschitz regularity constant of the matrix field A_b.

4.4.1 Viscosity regime

We construct a solution to the scheme (4.47) far enough from the point source singularity, at points $x \in \overline{\Omega}_h$ such that $|x| \geq r$, where r is independent of ε and h, by using the results developed in §4.3. For that purpose, a radius $r > 0$ is fixed in the rest of this section, unless otherwise specified, and such that $\mathbb{B}(6r) \subset \Omega$. The *erosion* with radius r of the domain Ω, and its intersection with the grid, are defined as

$$\mathrm{int}(\Omega, r) := \{x \in \Omega;\, \mathbb{B}(x, r) \subset \Omega\}, \qquad \mathrm{int}_h(\Omega, r) := \mathrm{int}(\Omega, r) \cap h\mathbb{Z}^d.$$

Lemma 4.4.6. *For each $\varepsilon > 0$ and $h > 0$ let u_h^ε be the solution to*

$$\mathcal{L}_h^\varepsilon u = 0 \text{ on } \Omega_h \setminus \overline{\mathbb{B}}(r), \qquad u = 1 \text{ on } \mathbb{S}(r) \qquad u = \exp(-R_{\mathcal{F}}/\varepsilon) \text{ on } \partial\Omega. \tag{4.52}$$

Then for $(\varepsilon, h/\varepsilon)$ sufficiently small, and denoting $\mathbf{u}_h^\varepsilon := -\varepsilon \ln u_h^\varepsilon$, one has with $C = C(M_{\mathcal{F}})$

$$|\mathbf{u}_h^\varepsilon(x) - \mathrm{dist}_{\mathcal{F}}(0, x)| \leq Cr \qquad \text{on } \mathrm{int}_h(\Omega, r) \setminus \mathbb{B}(2r). \tag{4.53}$$

Proof. Applying Theorem 4.3.18 to the domain $\Omega \setminus \mathbb{B}(r)$ we obtain that $\mathbf{u}_h^\varepsilon$ converges uniformly over the relatively compact subset $\mathrm{int}(\Omega, r) \setminus \mathbb{B}(2r)$ as $(\varepsilon, h/\varepsilon) \to 0$, to the limit

$$\mathbf{u}(x) = \min\big\{ \min_{p \in \mathbb{S}(r)} \mathrm{dist}_{\mathcal{F}}(p, x),\, R_{\mathcal{F}} + \min_{q \in \partial\Omega} \mathrm{dist}_{\mathcal{F}}(q, x)\big\} = \min_{p \in \mathbb{S}(r)} \mathrm{dist}_{\mathcal{F}}(p, x),$$

where the second equality follows from (4.50, right). Observing that $|\,\mathrm{dist}_{\mathcal{F}}(p, x) - \mathrm{dist}_{\mathcal{F}}(0, x)| \leq C_{\mathcal{F}}|p| \leq C_{\mathcal{F}}\, r$ for all $p \in \mathbb{S}(r)$, see (4.50, left), we conclude the proof. □

Corollary 4.4.7. *For $(\varepsilon, h/\varepsilon)$ sufficiently small, there exists $\underline{u}_h^\varepsilon : \overline{\Omega}_h \to \mathbb{R}$ such that $\tilde{\mathcal{L}}_h^\varepsilon \underline{u}_h^\varepsilon \geq 0$ and $\underline{\mathbf{u}}_h^\varepsilon(x) := -\varepsilon \ln \underline{u}_h^\varepsilon(x) \geq \mathrm{dist}_{\mathcal{F}}(0, x) - Cr$ on $\mathrm{int}_h(\Omega, r)$, where $C = C(M_{\mathcal{F}})$.*

Proof. From Lemma 4.4.6 introduce $\mathbf{u}_h^\varepsilon = -\varepsilon \ln u_h^\varepsilon$ obeying (4.53) for sufficiently small $(\varepsilon, h/\varepsilon)$, with constant $C_0 = C_0(M_{\mathcal{F}})$. Then let

$$\underline{u}_h^\varepsilon(x) := \begin{cases} 1 & x \in \mathbb{B}_h(2r), \\ \min\{1, u_h^\varepsilon(x) \exp(C_1 r/\varepsilon)\} & x \notin \overline{\Omega}_h \setminus \mathbb{B}_h(2r), \end{cases} \tag{4.54}$$

where $C_1 = C_0 + 3C_{\mathcal{F}}$. By construction one has $\underline{u}_h^\varepsilon(0) = 1$, and $\underline{u}_h^\varepsilon(x) \geq 0$ on $\partial\Omega$, so that $\tilde{\mathcal{L}}_h^\varepsilon u_h^\varepsilon \geq 0$ at these boundary points. By choice of the constant C_1 and in view of (4.53), one has $1 \leq u_h^\varepsilon(x) \exp(Cr/\varepsilon)$ on $\mathbb{B}_h(3r) \setminus \mathbb{B}_h(2r)$. Note that provided $h \leq r/R_{\mathcal{S}}$ the expression of $\mathcal{L}_h^\varepsilon \underline{u}_h^\varepsilon(x)$ at any $x \in \overline{\Omega}_h \setminus \mathbb{B}(3r)$ only involves values of $\underline{u}_h^\varepsilon$ in $\overline{\Omega}_h \setminus \mathbb{B}(2r)$. By Lemma 4.4.3, and since the constant 1 is a super-solution to $\mathcal{L}_h^\varepsilon$, we obtain that $\mathcal{L}_h^\varepsilon \underline{u}_h^\varepsilon \geq 0$, as announced. Finally, one has $\underline{\mathbf{u}}_h^\varepsilon(x) \geq \mathbf{u}_h^\varepsilon(x) - C_1 r \geq \mathrm{dist}_{\mathcal{F}}(0, x) - (C_0 + C_1)r$ on $\mathrm{int}_h(\Omega, r) \setminus \mathbb{B}_h(2r)$, and $\underline{\mathbf{u}}_h^\varepsilon(x) \geq 0 \geq \mathrm{dist}_{\mathcal{F}}(0, x) - 2C_{\mathcal{F}} r$ on $\mathbb{B}_h(2r)$, which concludes the proof. □

4.4.2 Taylor expansion regime

We construct explicit sub-solutions to the scheme (4.47), at points $h \lesssim |x| \lesssim \varepsilon$ and $\varepsilon \lesssim |x| \lesssim r$, which are radial functions with respectively a power and exponential profile. For that purpose, we need to estimate the derivatives of such functions.

Lemma 4.4.8. *Let $f \in C^2(\mathbb{R}_{++}, \mathbb{R})$, let $\mu \in \mathbb{R}$, and let $u(x) := \exp(-\mu f(|x|))$ for all $x \in \mathbb{R}^d \setminus \{0\}$. Then one has with $n(x) := x/|x|$, omitting the arguments of f, f', f'', f''' and n*

$$\frac{\nabla u(x)}{u(x)} = -\mu f' n, \qquad \frac{\nabla^2 u(x)}{u(x)} = \mu^2 f'^2 n n^\top + \mathcal{O}\big(\mu |f''| + \frac{\mu |f'|}{|x|}\big),$$

$$\frac{\nabla^3 u(x)}{u(x)} = \mathcal{O}\big(\mu^3|f'|^3 + \mu^2|f'||f''| + \frac{\mu^2|f'|^2}{|x|} + \frac{\mu|f''|}{|x|} + \mu|f'''| + \frac{\mu|f'|}{|x|^2}\big),$$

with absolute constants underlying the $\mathcal{O}$ notation.

Proof. The expression of $\nabla u(x)$ follows from the standard rules for the differentiation of an exponential function $\nabla(\exp\circ g) = (\exp\circ g)\nabla g$, and of a radial function $\nabla g(|x|) = g'(|x|)n(x)$. The full expression of $u^{-1}\nabla^2 u(x) = \mu^2 f'^2 nn^\top - \mu f'' nn^\top - \mu f'(\mathrm{Id} - nn^\top)/|x|$ can be obtained using the Leibniz rule for the differentiation of a product, and recalling that the Jacobian matrix of $n(x)$ is $(\mathrm{Id} - nn^\top)/|x|$. The expression of $\nabla^3 u$ follows likewise. □

Corollary 4.4.9. *Define $u(x) := \exp(-\lambda|x|/\varepsilon)$ where $\lambda \geq 1$, $\varepsilon > 0$. If $x \in \Omega_h$, $\varepsilon \leq |x| \leq 5r$ and $2R_{\mathcal{S}}h \leq \varepsilon$ then*

$$u(x)^{-1}\mathcal{L}_h^\varepsilon u(x) \leq 1 - \lambda^2|n(x)|^2_{A_b(x)} + C_0(\lambda + \lambda^3 h/\varepsilon). \tag{4.55}$$

In particular, $\mathcal{L}_h^\varepsilon u(x) \leq 0$ if $\lambda \geq C_1$ and $\lambda h/\varepsilon \leq c_2$, where $C_0, C_1, c_2 > 0$ only depend on $M_{\mathcal{F}}$.

Proof. Applying Lemma 4.4.8 to the identity function $f : r \in \mathbb{R}_{++} \mapsto r$, and parameter $\mu := \lambda/\varepsilon$ (note that $\mu \geq 1/\varepsilon$), we obtain whenever $|x| \geq \varepsilon/2$

$$\frac{\nabla u(x)}{u(x)} = \mathcal{O}(\mu), \qquad \frac{\nabla^2 u(x)}{u(x)} = \mu^2 nn^\top + \mathcal{O}(\frac{\mu}{\varepsilon}), \qquad \frac{\nabla^3 u(x)}{u(x)} = \mathcal{O}(\mu^3).$$

If $|x| \geq \varepsilon$ and $|e| \leq R_{\mathcal{S}}$, then any $y \in [x - he, x + he]$ obeys $|y| \geq \varepsilon/2$. Therefore

$$\frac{\overline{\delta}_h^e u(x)}{u(x)} = \mathcal{O}(\mu R_{\mathcal{S}} + h\mu^2 R_{\mathcal{S}}^2), \qquad \frac{\Delta_h^e u(x)}{u(x)} = \mu^2\langle n, e\rangle^2 + \mathcal{O}(\frac{\mu}{\varepsilon}R_{\mathcal{S}}^2 + h\mu^3 R_{\mathcal{S}}^3),$$

with again absolute constants underlying the $\mathcal{O}$ notation. Inserting these estimates in the scheme expression we obtain omitting the argument of ρ_i, $A_b^{-1}b$ and n

$$\frac{\mathcal{L}_h^\varepsilon u(x)}{u(x)} \leq 1 + 2\varepsilon C \sum_{1\leq i\leq I} \rho_i |\langle A_b^{-1}b, e_i\rangle|(\mu + h\mu^2) + \varepsilon^2 \sum_{1\leq i\leq I} \rho_i[-\mu^2\langle n, e_i\rangle^2 + C(\frac{\mu^2}{\varepsilon} + h\mu^3)],$$

where C depends only on $R_{\mathcal{S}}$. This establishes (4.55) observing that $\sum_{i=1}^I \rho_i\langle n, e_i\rangle^2 = \mathrm{Tr}(A_b nn^\top) = |n|^2_{A_b}$, and that $\sum_{i=1}^I \rho_i \leq \|\mathrm{Tr}(A_b)\|_\infty$. Since A_b is uniformly positive definite over $\overline{\Omega}$ and n is a unit vector, one has $|n|^2_{A_b} \geq c_0 = c_0(M_{\mathcal{F}}) > 0$, and the result follows with $C_1 = \max\{4C_0/c_0, \sqrt{2/c_0}\}$ and $c_2 = c_0/4C_0$. □

Corollary 4.4.10. *Define $u(x) := |x|^{-\mu}$, where $\mu \geq 1$. If $x \in \Omega_h$ and $2R_{\mathcal{S}}h \leq |x| \leq 4\varepsilon$ then*

$$\frac{\mathcal{L}_h^\varepsilon u(x)}{u(x)} \leq 1 - \frac{\varepsilon^2\mu^2}{|x|^2}|n(x)|^2_{A_b(x)} + C_0\big(\frac{\varepsilon^2\mu}{|x|^2} + \frac{h\varepsilon^2\mu^3}{|x|^3}\big). \tag{4.56}$$

In particular $\mathcal{L}_h^\varepsilon u(x) \leq 0$ if $\mu \geq C_1$ and $\mu h/\varepsilon \leq c_2$, where $C_0, C_1, c_2 > 0$ only depend on $M_{\mathcal{F}}$.

Proof. We apply Lemma 4.4.8 to the logarithm function $f = \ln$, obtaining

$$\frac{\nabla u(y)}{u(y)} = \mathcal{O}(\frac{\mu}{|y|}), \qquad \frac{\nabla^2 u(y)}{u(y)} = \frac{\mu^2 nn^\top}{|y|^2} + \mathcal{O}(\frac{\mu}{|y|^2}), \qquad \frac{\nabla^3 u(y)}{u(y)} = \mathcal{O}(\frac{\mu^3}{|y|^3}).$$

If $|x| \geq 2R_{\mathcal{S}}h$ and $|e| \leq R_{\mathcal{S}}$, then any $y \in [x - he, x + he]$ obeys $|y| \geq |x|/2$. Therefore

$$\frac{\overline{\delta}_h^e u(x)}{u(x)} = \mathcal{O}\big(\frac{\mu}{|x|} + \frac{h\mu^2}{|x|^2}\big), \qquad \frac{\Delta_h^e u(x)}{u(x)} = \frac{\mu^2\langle n, e\rangle^2}{|x|^2} + \mathcal{O}\big(\frac{\mu}{|x|^2} + \frac{h\mu^3}{|x|^3}\big).$$

Inserting these estimates in the scheme expression (4.32), we conclude similarly to Corollary 4.4.9. □

4.4.3 Finite neighborhood regime

We produce a sub-solution to the scheme $\tilde{\mathcal{L}}_h^\varepsilon$ which is useful in the immediate neighborhood of the origin, where $|x| \lesssim h$. The construction is not based on the approach of viscosity solutions, or on a Taylor expansion, but on the discrete structure of the scheme. For that purpose, we establish additional properties of its coefficients (4.51), suitably normalized: the first d offsets form a basis of $\mathbb{Z}^d$, and the corresponding weights are bounded below in a neighborhood of the source point. This implies that the stencils of our numerical scheme are locally connected, and allows to construct a subsolution in Corollary 4.4.13. The proof is based on the *spanning property* of Selling's decomposition, see Proposition 4.B.8, which is used here for the first time in the context of PDE numerical analysis.

Proposition 4.4.11. *Up to reordering the terms $(\rho_i, e_i)_{i=1}^I$ of Selling's decomposition (4.51) of the matrix field A_b, and grouping duplicate and opposite offsets $(e_i)_{i=1}^I$, one has for all $|x| \le r_{\mathcal{S}}$*

$$\min\{\rho_1(x), \cdots, \rho_d(x)\} \ge \rho_{\mathcal{S}}, \qquad \det(e_1, \cdots, e_d) = 1, \tag{4.57}$$

where the constants $\rho_{\mathcal{S}} > 0$ and $r_{\mathcal{S}} > 0$ only depend on $M_{\mathcal{F}}$.

Proof. Up to grouping duplicates and opposites, we may assume that the vectors $\pm e_1, \cdots, \pm e_I$ are pairwise distinct. Thus by Proposition 4.B.5 one has for all $x, y \in \overline{\Omega}$ and all $1 \le i \le I$

$$|\rho_i(x) - \rho_i(y)| \le C|x - y|, \tag{4.58}$$

where $C = C(M_{\mathcal{F}})$. Then by Proposition 4.B.8, and up to reordering $(\rho_i, e_i)_{i=1}^I$, one has $\det(e_1, \cdots, e_d) = 1$ and $\rho_i(0) \ge 2\rho_{\mathcal{S}}$ for all $1 \le i \le d$, where $\rho_{\mathcal{S}}$ only depends on $\|A_b(0)\|$ and $\|A_b(0)^{-1}\|$. The announced result follows, by choosing $r_{\mathcal{S}} := \rho_{\mathcal{S}}/C$. □

In the rest of this section, we assume that $(\rho_i, e_i)_{i=1}^I$ are ordered in such way that (4.57) holds. We also denote $\rho_{-i} := \rho_i$ and $e_{-i} := -e_i$ for all $1 \le i \le I$. Hence for any $x \in \Omega_h$ such that $\mathbb{B}(x, R_{\mathcal{S}} h) \subset \Omega_h$

$$\mathcal{L}_h^\varepsilon u(x) = \alpha_h^\varepsilon(x) u(x) - \sum_{1 \le |i| \le I} \beta_{h,i}^\varepsilon(x) u(x + h e_i),$$

where the coefficients are

$$\alpha_h^\varepsilon(x) := 1 + 2\frac{\varepsilon^2}{h^2} \sum_{1 \le i \le I} \rho_i(x), \qquad \beta_{h,i}^\varepsilon(x) := \rho_i(x)\Big(\frac{\varepsilon^2}{h^2} - \frac{\varepsilon}{h}\langle A_b(x)^{-1} b(x), e_i\rangle\Big). \tag{4.59}$$

Note that $\alpha_h^\varepsilon(x) \le 1 + 2(\varepsilon/h)^2 \|\operatorname{Tr}(A_b)\|_\infty$, since $\sum_{i=1}^I \rho_i(x) \le \sum_{i=1}^I \rho_i(x)|e_i|^2 = \operatorname{Tr}(A_b(x))$. We denote by $|x|_1$ the sum of the absolute values of the coefficients of a vector $x \in \mathbb{R}^d$.

Lemma 4.4.12. *Let $G \in \mathrm{GL}(\mathbb{Z}^d)$ be the matrix of columns $e_1, \cdots, e_d$, and let $N(x) := |G^{-1}x|_1$. Then for any $z \in \mathbb{Z}^d \setminus \{0\}$ there exists $1 \le |i| \le d$ such that $N(z + e_i) = N(z) - 1$. In addition $c|x| \le N(x) \le C|x|$ where the constants $C, c > 0$ only depend on $M_{\mathcal{F}}$.*

Proof. The matrix G has integer coefficients by construction, and $\det(G) = 1$ by (4.57, left) hence its inverse is the adjugate matrix $G^{-1} = \mathrm{co}(G)^\top$ which also has integer coefficients, thus $G \in \mathrm{GL}(\mathbb{Z}^d)$ as announced. Since the coefficients of G are bounded by $R_{\mathcal{S}}$, those of the adjugate matrix G^{-1} are bounded by $(d-1)! R_{\mathcal{S}}^{d-1}$, and the equivalence of N with the Euclidean norm follows.

Let $z \in \mathbb{Z}^d \setminus \{0\}$, and let $\lambda_1, \cdots, \lambda_d \in \mathbb{Z}$ be the coordinates of z in the basis $e_1, \cdots, e_d$, in other words $(\lambda_1, \cdots, \lambda_d)^\top = G^{-1} z$. Since $z \neq 0$, one at least of these coordinates is non-zero.

We thus assume w.l.o.g. that $\lambda_1 > 0$, up to a change of sign and permutation of the axes. Then $N(z - e_1) = |\lambda_1 - 1| + |\lambda_2| + \cdots + |\lambda_d| = -1 + |\lambda_1| + \cdots + |\lambda_d| = N(z) - 1$, which concludes the proof. □

Corollary 4.4.13. *Define $u(x) := \exp(-\nu N(x)/h)$. Then $\tilde{\mathcal{L}}_h^\varepsilon u(x) \leq 0$ on $\mathbb{B}_h(r_\mathcal{S})$, provided $\nu \geq \nu_0 = \nu_0(M_\mathcal{F})$, $\mathbb{B}(x, R_\mathcal{S} h) \subset \Omega$, and h/ε is sufficiently small.*

Proof. Note that $\beta_{h,i}^\varepsilon(x) \geq \rho_i(x)\varepsilon^2/(2h^2) \geq 0$, for all $1 \leq i \leq I$, when

$$h/\varepsilon \leq c := 1/(2\|A_b^{-1} b\|_\infty R_\mathcal{S}).$$

In particular $\beta_{h,i}^\varepsilon(x) \geq \rho_\mathcal{S}\varepsilon^2/(2h^2)$ if $|x| \leq r_\mathcal{S}$ and $1 \leq |i| \leq d$. By Lemma 4.4.12 there exists $1 \leq |i| \leq d$ such that $N(x + he_i) = N(x) - h$, and therefore $u(x + he_i) \geq e^\nu u(x)$. Thus

$$\frac{\mathcal{L}_h^\varepsilon u(x)}{u(x)} \leq \alpha_h^\varepsilon(x) - \beta_{h,i}^\varepsilon(x)\frac{u(x + he_i)}{u(x)} \leq 1 + 2\|\operatorname{Tr}(A_b)\|_\infty \frac{\varepsilon^2}{h^2} - e^\nu \frac{\rho_\mathcal{S}}{2}\frac{\varepsilon^2}{h^2}. \tag{4.60}$$

The result follows, by assuming in addition that $h \leq \varepsilon$ and choosing ν_0 such that $e^{\nu_0} := 2(1 + 2\|\operatorname{Tr}(A_b)\|_\infty)/\rho_\mathcal{S}$. □

4.4.4 Gluing the sub-solutions

In the previous subsections, we have produced four sub-solutions to the operator $\tilde{\mathcal{L}}_h^\varepsilon$, on different subsets of the domain $\overline{\Omega}_h$ defined according to the distance to the origin, see Lemma 4.4.6 and Corollaries 4.4.9, 4.4.10, and 4.4.13. We glue here these partial sub-solutions using Lemma 4.4.3, to produce a global sub-solution on $\overline{\Omega}_h$ and conclude the proof of Theorem 4.4.1. For that purpose, we introduce four mappings $u_h^{\varepsilon,i}$ defined on adequate subdomains $\Omega_h^{\varepsilon,i} \subset \overline{\Omega}_h$, $1 \leq i \leq 4$, and depending on the scale parameters (ε, h) as well as constants (λ, μ, ν, ξ) specified later.

- $u_h^{\varepsilon,0}(x) := v_h^\varepsilon(x) - \exp(-R_\mathcal{F}/\varepsilon)$, and $\Omega_h^{\varepsilon,0} := \overline{\Omega}_h \setminus \mathbb{B}_h(2r)$, where v_h^ε solves (4.52).
- $u_h^{\varepsilon,1}(x) = \exp(-\lambda|x|/\varepsilon)$, and $\Omega_h^{\varepsilon,1} := \mathbb{B}_h(5r) \setminus \mathbb{B}_h(\varepsilon)$.
- $u_h^{\varepsilon,2}(x) = |x|^{-\mu}$, and $\Omega_h^{\varepsilon,2} = \mathbb{B}_h(4\varepsilon) \setminus \mathbb{B}_h(2R_\mathcal{S} h)$.
- $u_h^{\varepsilon,3}(x) = \exp(-\nu N(x)/h)$, and $\Omega_h^{\varepsilon,3} = \mathbb{B}_h(\xi h)$, where N is from Lemma 4.4.12.

Proposition 4.4.14. *For any $(\varepsilon, h/\varepsilon)$ sufficiently small one has $\tilde{\mathcal{L}}_h^\varepsilon \overline{u}_h^\varepsilon \leq 0$ on $\overline{\Omega}_h$, where*

$$\overline{u}_h^\varepsilon(x) := \max\{u_h^{\varepsilon,3}(x),\, \alpha_2 h^\mu\, u_h^{\varepsilon,2}(x),\, \alpha_1(\tfrac{h}{\varepsilon})^\mu\, u_h^{\varepsilon,1}(x),\, \alpha_0(\tfrac{h}{\varepsilon})^\mu e^{-3\lambda\frac{r}{\varepsilon}}\, u_h^{\varepsilon,0}(x)\}, \tag{4.61}$$

for all $x \in \overline{\Omega}_h$, and where the quantity $u_h^{\varepsilon,i}(x)$ is only considered in the maximum if $x \in \Omega_h^{\varepsilon,i}$. The constants $(\lambda, \mu, \nu, \xi, \alpha_0, \alpha_1, \alpha_2)$ only depend on $M_\mathcal{F}$.

Proof. By Corollaries 4.4.9, 4.4.10, and 4.4.13 one may choose the constants λ, μ, ν such that $\tilde{\mathcal{L}}_\varepsilon^h u_h^{\varepsilon,i} \leq 0$ on $\Omega_h^{\varepsilon,i}$ for all $1 \leq i \leq 3$ and $(\varepsilon, h/\varepsilon)$ sufficiently small. Furthermore, this property is preserved if λ, μ or ν is increased. Also $\tilde{\mathcal{L}}_\varepsilon^h u_h^{\varepsilon,0} \leq 0$ on $\Omega_h^{\varepsilon,0}$, by noting that the positive constant $\exp(-R_\mathcal{F}/\varepsilon)$ subtracted in its definition accounts for the null boundary conditions of $\tilde{\mathcal{L}}_h^\varepsilon$, compare (4.47) with (4.52). Since the operator $\tilde{\mathcal{L}}_h^\varepsilon$ is linear on $\overline{\Omega}_h \setminus \{0\}$, see (4.47), the product of a sub-solution with a positive constant remains a sub-solution (outside the origin). Hence (4.61) is a maximum of 4 sub-solutions on their respective domains.

We next proceed to prove estimates of the following form: for any $x \in \Omega_h^{\varepsilon,i} \cap \Omega_h^{\varepsilon,i+1}$

$$m_h^{\varepsilon,i} u_h^{\varepsilon,i}(x) \leq (\text{resp. } \geq)\, u_h^{\varepsilon,i+1}(x) \quad \text{when} \quad \mathbb{B}_h(x, R_{\mathcal{S}} h) \not\subset \Omega_h^{\varepsilon,i} \text{ (resp. } \Omega_h^{\varepsilon,i+1}), \tag{4.62}$$

where $m_h^{\varepsilon,i}$ is a suitable function of the scale parameters. Thus by Lemma 4.4.3,

$$u_h^\varepsilon(x) := \max\{u_h^{3,\varepsilon}(x),\ m_h^{\varepsilon,2} u_h^{2,\varepsilon}(x),\ m_h^{\varepsilon,2} m_h^{\varepsilon,1} u_h^{1,\varepsilon}(x),\ m_h^{\varepsilon,2} m_h^{\varepsilon,1} m_h^{\varepsilon,0} u_h^{1,\varepsilon}(x)\}$$

is a sub-solution, which is the announced result. Indeed one has $\mathcal{L}_h^\varepsilon u_h^{\varepsilon,i}(x) \leq 0$ if $\mathbb{B}_h(x, R_{\mathcal{S}} h) \subset \Omega_h^{\varepsilon,i}$, but $\mathcal{L}_h^\varepsilon u_h^{\varepsilon,i}(x)$ may not make sense if $\mathbb{B}_h(x, R_{\mathcal{S}} h) \not\subset \Omega_h^{\varepsilon,i}$ since it could involve values of $u_h^{\varepsilon,i}$ outside $\Omega_h^{\varepsilon,i}$; in that case however, (4.62) shows that $\overline{u}_h^\varepsilon(x)$ is not defined from $u_h^{\varepsilon,i}(x)$.

The estimates (4.62) follow from basic upper and lower bounds of the involved functions, and of the norms of the relevant points x. Namely

$$\begin{aligned} &u_h^{\varepsilon,0}(x) \leq 1, && u_h^{\varepsilon,1}(x) \geq \exp(-3\lambda r/\varepsilon), && \text{when } 2r \leq |x| \leq 3r. \\ &u_h^{\varepsilon,0}(x) \geq \exp(-Cr/\varepsilon), && u_h^{\varepsilon,1}(x) \leq \exp(-4\lambda r/\varepsilon), && \text{when } 4r \leq |x| \leq 5r. \end{aligned}$$

The upper bound on $u_h^{\varepsilon,0}$ is derived from the maximum principle, and the lower bound from Lemma 4.4.6, with $C = C(M_{\mathcal{F}})$ and for sufficiently small $(\varepsilon, h/\varepsilon)$. This establishes (4.62, $i = 0$) with $m_h^{\varepsilon,0} = \exp(-3\lambda r/\varepsilon)$, up to increasing λ so that $\lambda \geq C$. Likewise

$$\begin{aligned} &u_h^{\varepsilon,1}(x) \leq \exp(-\lambda), && u_h^{\varepsilon,2}(x) \geq (2\varepsilon)^{-\mu}, && \text{when } \varepsilon \leq |x| \leq 2\varepsilon. \\ &u_h^{\varepsilon,1}(x) \geq \exp(-4\lambda), && u_h^{\varepsilon,2}(x) \leq (3\varepsilon)^{-\mu}, && \text{when } 3\varepsilon \leq |x| \leq 4\varepsilon. \end{aligned}$$

This establishes (4.62, $i = 1$) with $m_h^{\varepsilon,1} = e^\lambda (2\varepsilon)^{-\mu}$, up to increasing μ so that $(3/2)^\mu \geq e^{3\lambda}$. Lastly

$$\begin{aligned} &u_h^{\varepsilon,2}(x) \leq (2R_{\mathcal{S}} h)^{-\mu}, && u_h^{\varepsilon,3}(x) \geq \exp(-3R_{\mathcal{S}} C_N \nu), && \text{when } 2R_{\mathcal{S}} h \leq |x| \leq 3R_{\mathcal{S}} h. \\ &u_h^{\varepsilon,2}(x) \geq (\xi R_{\mathcal{S}} h)^{-\mu}, && u_h^{\varepsilon,3}(x) \leq \exp(-(\xi - R_{\mathcal{S}}) c_N \nu), && \text{when } (\xi - R_{\mathcal{S}}) h \leq |x| \leq \xi R_{\mathcal{S}} h, \end{aligned}$$

where c_N and C_N are the equivalence constants in Lemma 4.4.12. We define ξ by $(\xi - R_{\mathcal{S}}) c_N - 3R_{\mathcal{S}} C_N = 1$. This establishes (4.62, $i = 2$) with $m_h^{\varepsilon,2} = e^{-3R_{\mathcal{S}} C_N \mu} (2R_{\mathcal{S}} h)^\mu$, up to increasing ν so that $e^\nu \geq (\xi/(2R_{\mathcal{S}}))^\mu$, in view of the expression of ξ, which concludes the proof. □

Corollary 4.4.15. *For $(\varepsilon, h/\varepsilon)$ sufficiently small, there exists $\overline{u}_h^\varepsilon : \overline{\Omega}_h \to \mathbb{R}$ such that $\tilde{\mathcal{L}}_h^\varepsilon \overline{u}_h^\varepsilon \leq 0$ and $\underline{\mathbf{u}}_h^\varepsilon(x) := -\varepsilon \ln \overline{u}_h^\varepsilon(x) \leq \mathrm{dist}_{\mathcal{F}}(0, x) + C(r + \varepsilon \ln(\varepsilon/h))$ on $\mathrm{int}_h(\Omega, r)$, where $C = C(M_{\mathcal{F}})$.*

Proof. We distinguish two cases. (i) If the maximum in (4.61) is attained by the last term, then the announced result follows Lemma 4.4.6 and the expression of the multiplicative factor $\alpha_0 (h/\varepsilon)^\mu \exp(-3\lambda r/\varepsilon)$. (ii) If the maximum in (4.61) is attained by one of the first three terms, then $|x| \leq 5r$ and the announced result follows from the explicit expressions of $u_h^{\varepsilon,1}, u_h^{\varepsilon,2}, u_h^{\varepsilon,3}$ as well as $\mathrm{dist}_{\mathcal{F}}(0, x) \leq 5C_{\mathcal{F}} r$. □

Proof of Theorem 4.4.1. For sufficiently small $(\varepsilon, h/\varepsilon)$, we obtain from the comparison principle Lemma 4.3.4 and with the mappings $\overline{\mathbf{u}}_h^\varepsilon$ and $\underline{\mathbf{u}}_h^\varepsilon$ of Corollaries 4.4.7 and 4.4.15 respectively that

$$\mathrm{dist}_{\mathcal{F}}(0, x) - Cr \leq \overline{\mathbf{u}}_h^\varepsilon(x) \leq \mathbf{u}_h^\varepsilon(x) \leq \underline{\mathbf{u}}_h^\varepsilon(x) \leq \mathrm{dist}_{\mathcal{F}}(0, x) + C(r + \varepsilon \ln(\varepsilon/h)), \tag{4.63}$$

on $\mathrm{int}_h(\Omega, r)$, where $C = C(M_{\mathcal{F}})$. Since the parameter $r > 0$ is arbitrary[2], except for the constraint $\mathbb{B}(6r) \subset \Omega$, we conclude as announced that $\mathbf{u}_h^\varepsilon(x) \to \mathrm{dist}_{\mathcal{F}}(0, x)$ locally uniformly on Ω as $(\varepsilon, h/\varepsilon, \varepsilon \ln(\varepsilon/h)) \to 0$. The result follows, noting that $\varepsilon \ln(\varepsilon/h) \leq \varepsilon |\ln h|$ when $0 < h \leq \varepsilon \leq 1$. □

[2]Note nevertheless that (4.63) holds when $\varepsilon \leq \delta$ and $h/\varepsilon \leq \delta$, where δ depends on $M_{\mathcal{F}}$ and r.

4.4.5 Convergence on $\Omega \times \Omega$ and inverse matrix

We establish Theorem 4.4.2, which relates the Randers distance with the inverse matrix of our finite differences scheme. For that purpose, we use the following convention: if $U(x;x_*)$ if a bivariate discrete mapping, defined for all $(x,x_*) \in \overline{\Omega}_h \times \Omega_h$, and if F is a finite differences scheme of the form of Definition 4.3.3, then $FU(x;x_*) := \tilde{F}(x, U(x;x_*), [U(x;x_*) - U(y;x_*)]_{y \in X \setminus \{x\}})$. In other words, the numerical scheme sees U as a function of its first variable x only.

Lemma 4.4.16. *For any $(\varepsilon, h/\varepsilon)$ sufficiently small, and any $x_* \in \mathbb{B}_h(r/2)$, one has $\tilde{\mathcal{L}}_h^\varepsilon U_h^\varepsilon(x;x_*) \leq 0$ on $\Omega_h \setminus \{x_*\}$, where for all $x \in \overline{\Omega}_h$*

$$\overline{U}_h^\varepsilon(x;x_*) := \max\{u_h^{\varepsilon,3}(x),\, \alpha_2 h^\mu\, u_h^{\varepsilon,2}(x-x_*),\, \alpha_1 (h/\varepsilon)^\mu\, u_h^{\varepsilon,1}(x-x_*),\\ \alpha_0 (h/\varepsilon)^\mu e^{-3\lambda r/\varepsilon}\, u_h^{\varepsilon,0}(x-x_*)\},$$

and where the quantity $u_h^{\varepsilon,i}(x-x_*)$ is only considered in the maximum if $x - x_* \in \Omega_h^{\varepsilon,i}$. The constants $(\lambda,\mu,\nu,\xi,\alpha_0,\alpha_1,\alpha_2)$ only depend on $M_{\mathcal{F}}$. In addition $\underline{\mathbf{U}}_h^\varepsilon(x;x_*) := -\varepsilon \ln \overline{U}_h^\varepsilon(x;x_*) \leq \mathrm{dist}_{\mathcal{F}}(0,x) + C(r + \varepsilon \ln(\varepsilon/h))$ for all $(x,x_*) \in \mathrm{int}_h(\Omega,r) \times \mathbb{B}_h(r/2)$, where $C = C(M_{\mathcal{F}})$.

Proof. The proofs of Proposition 4.4.14 and Corollary 4.4.7 adapt in a straightforward manner to a point source x_* sufficiently close to the origin, as here. □

Proposition 4.4.17 (Convergence in the product space)**.** *Denote by $U_h^\varepsilon : \overline{\Omega}_h \times \Omega_h \to \mathbb{R}$ the solution to*

$$\mathcal{L}_h^\varepsilon U_h^\varepsilon(x;\, x_*) = 0, \forall x \in \Omega_h \setminus \{x_*\}, \qquad U_h^\varepsilon(x_*;\, x_*) = 1 \qquad U_h^\varepsilon(x;x_*) = 0, \forall x \in \partial\Omega. \tag{4.64}$$

Then locally uniformly on $\Omega \times \Omega$ one has $-\varepsilon \ln U_h^\varepsilon(x;x_*) \to \mathrm{dist}_{\mathcal{F}}(x_*,x)$ as $(\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0$.

Proof. First note that $x \in \overline{\Omega}_h \mapsto U(x;x_*)$, for any given $x_* \in \Omega_h$, solves a linear problem which is elliptic when h/ε is sufficiently small, hence has a unique solution, see Corollary 4.3.5 and Proposition 4.3.9.

Let $r > 0$ be such that $\mathbb{B}(6r) \subset \Omega$. Then for $(\varepsilon, h/\varepsilon)$ sufficiently small and for all $(x,x_*) \in \mathrm{int}_h(\Omega,r) \times \mathbb{B}_h(r/2)$ one has by Corollary 4.4.7 and Lemma 4.4.16 and for some constant $C = C(M_{\mathcal{F}})$

$$\mathrm{dist}_{\mathcal{F}}(0,x) - Cr \leq \overline{\mathbf{U}}_h^\varepsilon(x;x_*) \leq \mathbf{U}_h^\varepsilon(x;x_*) \leq \underline{\mathbf{u}}_h^\varepsilon(x) \leq \mathrm{dist}_{\mathcal{F}}(0,x) + C(r + \varepsilon \ln(\tfrac{\varepsilon}{h})), \tag{4.65}$$

and therefore $|\mathbf{U}(x;x_*) - \mathrm{dist}_{\mathcal{F}}(x_*,x)| \leq (2C + C_{\mathcal{F}})r$ when in addition $\varepsilon \ln(\varepsilon/h) \leq r$, noting that $|\mathrm{dist}_{\mathcal{F}}(x_*,x) - \mathrm{dist}_{\mathcal{F}}(0,x)| \leq C_{\mathcal{F}} r$.

Now let $K_* \subset \Omega$ be a compact set. Up to reducing r one can find a finite cover $K_* \subset \cup_{j=1}^J \mathbb{B}(y_j, r/2)$ such that $\mathbb{B}(y_j, 6r) \subset \Omega$ for all $1 \leq j \leq J$. Applying the above reasoning to each ball $\mathbb{B}_h(y_j, r/2)$, $1 \leq j \leq J$, instead of $\mathbb{B}_h(r/2)$, we obtain $|\mathbf{U}(x;x_*) - \mathrm{dist}_{\mathcal{F}}(x_*,x)| \leq (2C + C_{\mathcal{F}})r$ for all $(x,x_*) \in \mathrm{int}_h(\Omega,r) \times (K_* \cap h\mathbb{Z}^d)$, when $(\varepsilon, h/\varepsilon, \varepsilon \ln h)$ is small enough. Since r can be chosen arbitrarily small, the result follows. □

Lemma 4.4.18. *If h/ε is sufficiently small, then for all $x_* \in \Omega_h$ such that $\mathbb{B}(x_*, R_{\mathcal{S}} h) \subset \Omega$ one has $1 \leq \mathcal{L}_h^\varepsilon U_h^\varepsilon(x_*;x_*) \leq 1 + C\frac{\varepsilon^2}{h^2}$ where $C = 2\|\mathrm{Tr}(A_b)\|_\infty$.*

Proof. We assume that $C_0 h \leq \varepsilon$ where $C_0 = \|A_b^{-1} b\|_\infty R_{\mathcal{S}}$, and obtain by Proposition 4.3.9 that $\mathcal{L}_h^\varepsilon$ is DDE. By the comparison principle, one has $0 \leq U_h^\varepsilon(x;x_*) \leq 1$ for all $x \in \Omega_h$. Thus $1 \leq \mathcal{L}_h^\varepsilon U_h^\varepsilon(x_*,x_*) \leq a_h^\varepsilon(x_*)$, with the notations (4.59), since $\beta_{h,i}^\varepsilon(x_*) \geq 0$ for all $1 \leq i \leq I$. The result follows. □

Proof of inverse matrix convergence, Theorem 4.4.2. By definition of L_h^ε and U_h^ε

$$(L_h^\varepsilon)^{-1}_{x_* x} = \frac{U_h^\varepsilon(x; x_*)}{\mathcal{L}_h^\varepsilon U_h^\varepsilon(x_*; x_*)}.$$

Thus $\varepsilon|\ln[(L_h^\varepsilon)^{-1}_{x_* x}] - \ln U_h^\varepsilon(x; x_*)| \leq \varepsilon \ln(1 + C\varepsilon^2/h^2)$, under the conditions of Lemma 4.4.18. Noting that $\varepsilon \ln(1 + C\varepsilon^2/h^2) \to 0$ as $(\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0$, and that $-\varepsilon \ln U_h^\varepsilon(x; x_*) \to \mathrm{dist}_{\mathcal{F}}(x_*, x)$ locally uniformly by Proposition 4.4.17, we conclude the proof. □

4.5 Application to regularized optimal transport

In this section, we describe a numerical approach to the 1-Wasserstein optimal transport problem, with cost defined as a Randers distance, and with entropic relaxation. Given probability measures $\mu, \nu \in \mathcal{P}(\Omega)$, the addressed problem reads

$$W_\varepsilon(\mu, \nu) := \inf_{P \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} C(x, y)\, \mathrm{d}P(x, y) - \varepsilon\, \mathrm{Ent}(P), \tag{4.66}$$

where $\varepsilon \geq 0$ is the entropic relaxation parameter, and where $\Pi(\mu, \nu)$ is the set of probability measures on $\Omega \times \Omega$ whose first and second marginals coincide respectively with μ and ν, known as *transport plans* between μ and ν. The transport cost and entropy are defined as

$$C(x, y) := \mathrm{dist}_{\mathcal{F}}(x, y), \qquad \mathrm{Ent}(P) := -\int_{\Omega\times\Omega} \ln\Big(\frac{\mathrm{d}P(x, y)}{e\, \mathrm{d}P_0(x, y)}\Big)\mathrm{d}P(x, y)$$

where $\mathcal{F}$ is a Randers metric on the domain Ω, subject to the well posedness assumptions listed in the last paragraph of section 4.1, and P_0 is a reference measure on $\Omega \times \Omega$. The Euler constant e appearing in $\mathrm{Ent}(P)$ only changes the entropy by an additive constant, since P has total mass one, and allows simplifying later calculations.

As mentioned in the introduction, our approach extends [Cut13] from Riemannian to non-symmetric Randers metrics. However, the quadratic cost $\mathrm{dist}_{\mathcal{F}}(x, y)^2$ corresponding to the 2-Wasserstein distance cannot be addressed in our setting, see Remark 4.4.5. Let us also acknowledge that the effect of entropic relaxation cannot be ignored in the numerical implementation of this class of methods: indeed, empirically, the transport plan is blurred over a radius $\sqrt{\varepsilon}$, while ε itself must be substantially larger than the discretization grid scale, see Theorem 4.3.18. Nevertheless such as smoothing is not necessarily an issue in applications [Cut13], and the estimation of the Wasserstein distance itself as $\varepsilon \to 0$ can be accelerated by suitable techniques [Chi+20].

4.5.1 Kantorovich duality, and Sinkhorn's algorithm

We assume in the following that μ and ν are supported on a finite set $X \subset \Omega$, and present in this setting Kantorovich's dual formulation of the optimal transport problem (4.66), and its numerical solution by Sinkhorn's algorithm. With a slight abuse of notation, we identify a measure μ on the finite set X (resp. P on $X \times X$), which is a weighted sum of Dirac masses $\mu = \sum_{x\in X} \mu_x \delta_x$, with the corresponding non-negative vector $(\mu_x)_{x\in X}$ (resp. matrix $(P_{xy})_{x,y\in X}$). With this convention, the set of probability measures on X, and of transport plans between two such probabilities, are defined as

$$\mathcal{P}(X) := \{\mu \in \mathbb{R}_+^X;\, \mu^\top \mathbb{1} = 1\}, \qquad \Pi(\mu, \nu) := \{P \in \mathbb{R}_+^{X\times X};\, P\mathbb{1} = \mu,\, P^\top \mathbb{1} = \nu\}, \tag{4.67}$$

where $\mathbb{R}_+ := [0, \infty[$ denotes the set of non-negative reals, and $\mathbb{1} = (1, \cdots, 1)^\top \in \mathbb{R}^X$. In particular, $\mu^\top \mathbb{1} = \sum_{x \in X} \mu_x$, $P\mathbb{1} = (\sum_{y \in X} P_{xy})_{x \in X}$, and $P^\top \mathbb{1} = (\sum_{x \in X} P_{xy})_{y \in X}$. In this discrete setting, the optimal transport problem (4.66) reads

$$W_\varepsilon(\mu, \nu) = \inf_{P \in \Pi(\mu,\nu)} \langle\!\langle P, C \rangle\!\rangle + \varepsilon \, \langle\!\langle P, \ln\big(\frac{P}{eP_0}\big) \rangle\!\rangle, \tag{4.68}$$

where $\langle\!\langle A, B \rangle\!\rangle := \mathrm{Tr}(A^\top B) = \sum_{x,y \in X} A_{xy} B_{xy}$. In (4.68) and below, the fraction bar, the logarithm and the exponential function apply componentwise to vectors and matrices. We assume that the reference measure $P_0 = (P^0_{xy})$ has positive entries, and use the standard convention $0 \times \infty = 0$ in the definition of the entropic term if some entries of $P \in \Pi(\mu, \nu)$ vanish. Noting that $s \in \mathbb{R}_{++} \mapsto s \ln s$ is convex and has a vertical tangent at the origin, we find that the minimization problem (4.68) is convex and that the optimal P has positive entries whenever $\varepsilon > 0$.

Kantorovich duality introduces potentials $\varphi, \psi \in \mathbb{R}^X$ to account for the equality constraints in (4.67), and uses Sion's minimax theorem [Kom88] to re-order the sup and inf:

$$\begin{aligned} W_\varepsilon(\mu, \nu) &= \inf_{P \in \mathbb{R}_+^{X \times X}} \Big(\langle\!\langle P, C \rangle\!\rangle + \varepsilon \, \langle\!\langle P, \ln\big(\frac{P}{eP_0}\big) \rangle\!\rangle + \sup_{\varphi, \psi \in \mathbb{R}^X} \langle \varphi, \mu - P\mathbb{1} \rangle + \langle \psi, \nu - P^\top \mathbb{1} \rangle \Big) \\ &= \sup_{\varphi, \psi \in \mathbb{R}^X} \Big(\langle \varphi, \mu \rangle + \langle \psi, \nu \rangle + \inf_{P \in \mathbb{R}_+^{X \times X}} \langle\!\langle P, \, C + \varepsilon \ln\big(\frac{P}{eP_0}\big) - \varphi \mathbb{1}^\top - \mathbb{1} \psi^\top \rangle\!\rangle \Big) \\ &= \sup_{\varphi, \psi \in \mathbb{R}^X} \langle \varphi, \mu \rangle + \langle \psi, \nu \rangle - \varepsilon \, \langle\!\langle P_0, \, \exp\big(\frac{\varphi \mathbb{1}^\top + \mathbb{1} \psi^\top - C}{\varepsilon}\big) \rangle\!\rangle. \end{aligned} \tag{4.69}$$

The third line was obtained by solving, component-wise and in closed form, the minimization w.r.t. P. Namely, the convex one dimensional mapping $p \in \mathbb{R}_{++} \mapsto p\big(C_{xy} + \varepsilon \ln\big[p/(eP^0_{xy})\big] - \varphi_x - \psi_y\big)$ attains its minimum for

$$P_{xy} = P^0_{xy} \exp[(\varphi_x + \psi_y - C_{xy})/\varepsilon]. \tag{4.70}$$

Using the change of variables $\Phi = \exp(\varphi/\varepsilon)$ and $\Psi := \exp(\psi/\varepsilon)$ we conclude that

$$W_\varepsilon(\mu, \nu) = \varepsilon \max_{\Phi, \Psi \in \mathbb{R}^X_{++}} \langle \ln \Phi, \mu \rangle + \langle \ln \Psi, \nu \rangle - \langle \Phi^\top, K_\varepsilon \Psi \rangle, \tag{4.71}$$

where $K_\varepsilon = (K^\varepsilon_{xy})_{x,y \in X}$ with $K^\varepsilon_{xy} := P^0_{xy} \exp(-C_{xy}/\varepsilon)$. Note that the maximization problem (4.69) is strictly concave. The equivalent form (4.71) can be numerically solved using alternate maximization, in other words successively solving w.r.t. the unknown Φ with Ψ fixed (resp. w.r.t. Ψ with Φ fixed). This approach is known as Sinkhorn's algorithm [Sin64], and is particularly simple and efficient since the optimal value w.r.t. either of these variables has a closed form, when the other variable is fixed. More precisely, given an arbitrary $\Psi_0 \in \mathbb{R}^X_{++}$ one defines for all $n \geq 0$

$$\Phi_n := \frac{\mu}{K_\varepsilon \Psi_n}, \qquad\qquad \Psi_{n+1} := \frac{\nu}{K_\varepsilon^\top \Phi_n}, \tag{4.72}$$

where, as in (4.68), the fraction bar denotes a componentwise division operation. Then the sequence $(\Phi_n, \Psi_n)_{n \geq 0}$ converges geometrically to a maximizer of (4.71), see [Sin64].

4.5.2 Efficient computation

The more computationally intensive part of Sinkhorn's algorithm (4.72) is to repeatedly compute the matrix-vector products $K_\varepsilon \Phi_n$ and $K_\varepsilon^\top \Psi_n$ in (4.72), since the matrix K_ε is dense and large. An efficient way to approximate those products using Varadhan's formula was proposed in [Sol+15],

in the case of Riemannian manifolds. We adapt here this approach to Randers manifolds, thus specializing to the case $K^\varepsilon_{xy} := \exp(-\operatorname{dist}_{\mathcal{F}}(x,y)/\varepsilon)$ where the reference measure $P_0 \equiv 1$ is the uniform probability, the transport cost is defined as a Randers distance $\operatorname{dist}_{\mathcal{F}}$, and where $X = \Omega_h$ is a domain discretized on a Cartesian grid of scale $h > 0$.

Under these conditions, denoting by L^ε_h the matrix of our linear discretization scheme (4.32) with null boundary conditions, one has by Theorem 4.4.2

$$[L^\varepsilon_h]^{-1}_{xy} = \exp(-\frac{\operatorname{dist}_{\mathcal{F}}(x,y) + o(1)}{\varepsilon}), \qquad \text{as } (\varepsilon, h/\varepsilon, \varepsilon \ln h) \to 0, \tag{4.73}$$

locally uniformly on $\Omega \times \Omega$. Therefore the dense matrix product $\Phi' = K_\varepsilon \Phi$ can be approximated by solving the sparse linear system $\Phi = L^\varepsilon_h \Phi'$, which is considerably less memory intensive, and has a lower complexity along the iterations especially if a sparse pre-factorization of the matrix L^ε_h is used.

4.6 Numerical results

We illustrate the numerical methods presented in this paper, for Randers distance computation and numerical optimal transport, with synthetic numerical experiments in dimension $d = 2$. Geodesic distance computation based on solving the heat or Poisson PDEs has already numerous applications [CWW13; YC16; Yan+18] and is part of established algorithmic geometry libraries such as CGAL®. Likewise Wasserstein distance computation based on entropic relaxation is an established numerical approach [Cut13; Sol+15; Chi+20]. The contributions of this paper are thus mostly theoretical, see section 4.7.

The approach presented in this paper for Randers distance computation is applied in [Yan+18] to image segmentation problems, using numerical codes provided by Jean-Marie Mirebeau and with due acknowledgement[3]. Optimal transport w.r.t. Randers geometry and the present numerical method is yet to find a concrete application, but let us nevertheless mention the following motivation which was recently presented to us: monitoring forest fires using a fleet of small drones, which requires spreading the agents over a large specified area, and involves strongly asymmetrical displacement costs depending on the winds and terrain.

In this numerical section, we compare in several occasions the results of the centered scheme $\mathcal{L}^\varepsilon_h$ (4.32) emphasized in this paper, with those of the upwind scheme $\mathcal{L}^{\varepsilon,+}_h$ (4.38) which is unconditionally stable but is also less accurate. We limit our experiments to two-dimensional problems, consistently with the literature, and although our theoretical results apply in dimension three as well, due to the overwhelming cost of solving three-dimensional Laplacian-like linear systems at the considered grid scales.

The PDE domain for the experiments presented in this section is the two-dimensional unit ball $\Omega = \{x \in \mathbb{R}^2; |x| \leq 1\}$, which is discretized on a regular Cartesian grid, using finite differences modified as in (4.28) to account for the (null) boundary conditions on $\partial\Omega$. The grid scale $h = 0.00625$ commonly used in the experiments below corresponds to a grid of size 320×320 (intersected with the ball). In the first two problems we numerically approximate

$$\mathbf{u}(x) := \min_{y \in Y} \operatorname{dist}_{\mathcal{F}}(x,y), \tag{4.74}$$

where Y is a finite set of target points, and $\mathcal{F}$ is a Randers metric on Ω which is described in terms of the parameters A, b of its dual, see Lemma 4.2.6. From the convergence analysis standpoint,

[3] However [Yan+18, §2.2] attempts to relate the numerical method with the Finsler heat equation (4.49). This is incorrect to our belief, and was published without the knowledge of the authors of this paper.

the case of finitely many isolated point sources is a straightforward generalization of the case of a single one considered section 4.4, and considering targets instead of sources amounts to a change of sign in the asymmetric part of the metric as discussed below (4.4).

In our experiments, the largest contributor to computation time is the factorization of the sparse linear systems, using the SuperLU routine provided with the scipy Python package. In contrast, the preliminary step of scheme construction (including Selling's algorithm to decompose the matrix $A_b(x)$ at each point $x \in \Omega_h$, and sparse matrix assembly) only accounts for fraction of this cost, and the subsequent solve operation is approximately $10\times$ faster than matrix factorization. In the application to optimal transport, which is based on Sinkhorn's algorithm (4.72), the same linear system needs to be solved multiple times, and thus a single matrix factorization is followed by 13 to 54 solve operations. The SuperLU factorization time when using a 320×320 discretization grid (thus $\approx 10^5$ unknowns) ranges from $1s$ to $1.6s$ depending on the test case, on a laptop equipped with a 2.3 GHz Intel Core i5 dual-core processor.

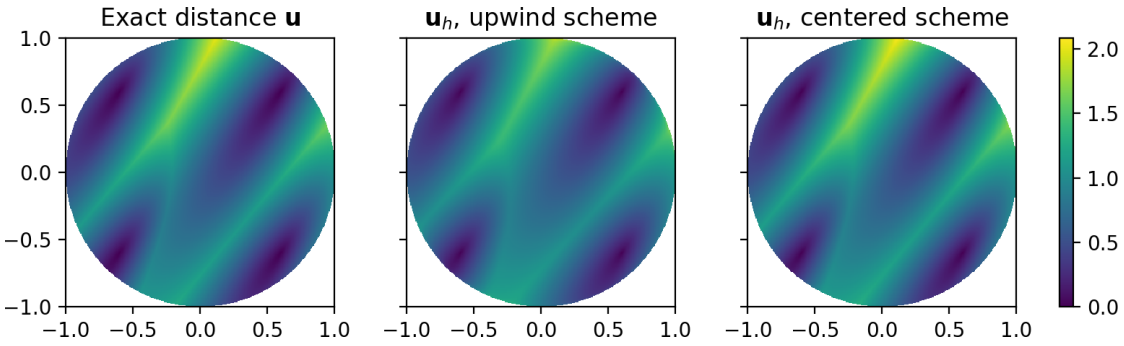


Figure 4.1: Randers distance with parameters (4.75). Left: exact solution. Center : solution based on the upwind scheme $\mathcal{L}_h^{\varepsilon,+}$ (4.38). Right : more accurate solution based on the centered scheme $\mathcal{L}_h^{\varepsilon}$ (4.32). In all cases $h = 0.00625$, $\varepsilon = 0.5h^{2/3}$.

Randers metric with constant coefficients We consider a finite set Y of target points and a Randers metric whose dual $\mathcal{F}^*$ is defined by the following coefficients A, b

$$A := \begin{pmatrix} 0.5 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}, \quad b := \begin{pmatrix} 0.3 \\ 0.4 \end{pmatrix}, \quad Y := \left\{ \begin{pmatrix} -0.6 \\ 0.6 \end{pmatrix}, \begin{pmatrix} -0.6 \\ -0.6 \end{pmatrix}, \begin{pmatrix} 0.6 \\ -0.6 \end{pmatrix}, \begin{pmatrix} 0.6 \\ 0.6 \end{pmatrix} \right\}. \tag{4.75}$$

Since the metric is constant and the domain is convex, the geodesic distance is explicit: $\mathrm{dist}_{\mathcal{F}}(x, y) = F(y - x)$ where $\mathcal{F}_x(v) = F(v)$ for all $x \in \Omega$, and the minimal paths are straight lines, see the discussion below Definition 4.2.3. In particular (4.74) can be evaluated exactly, which allows estimating convergence rates.

The exact Randers distance from Y, and its approximation produced using the centered scheme (4.32) and the upwind scheme (4.38), are illustrated on Figure 4.1. We present on Figure (4.2, top left) Tissot's indicatrix of the metric $\mathcal{F}$, which is a representation of the sets

$$\{x + v;\, v \in \mathbb{R}^2,\, \mathcal{F}_x(v) = r\}, \tag{4.76}$$

at a number of points $x \in \Omega$ and for a suitable radius $r > 0$. In Randers case, the set (4.76) is an ellipse which is *not* centered on the point x, and admits several equivalent characterizations see Lemma 4.2.7. The numerical approximation of Randers distance obtained with the centered scheme is illustrated on Figure (4.2, top right), while the numerical approximations of minimal paths from Y obtained by solving the ODE (4.17) are shown Figure (4.2, top center).

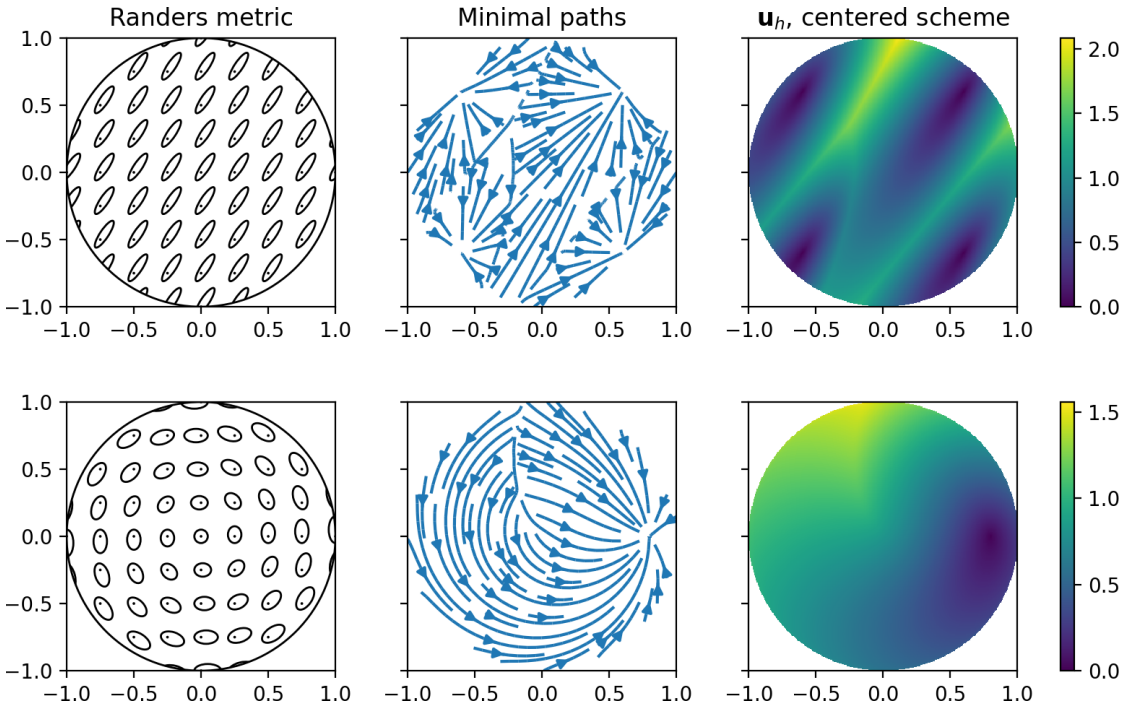


Figure 4.2: Representation of the Randers metric and approximations of minimal paths and of the Randers distance for parameters (4.75) (top), and (4.77) (bottom), with $h = 0.00625$ and $\varepsilon = 0.5h^{2/3}$.

Randers metric with variable coefficients A single target point is considered $Y = \{(0.8, 0)\}$, and the dual metric parameters are defined at $x = (x_1, x_2) \in \overline{\Omega}$ as

$$A(x) := \begin{pmatrix} 1 + \frac{2x_2^2}{|x|} + x_2^2 & -\frac{2x_1x_2}{|x|} - x_1x_2 \\ -\frac{2x_1x_2}{|x|} - x_1x_2 & 1 + \frac{2x_1^2}{|x|} + x_1^2 \end{pmatrix}, \qquad b(x) := x^\perp = \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, \tag{4.77}$$

where A is extended by continuity at the origin. Numerical results are shown Figure (4.2, bottom).

Numerical convergence rates We discuss the convergence of some approximations of the exact distance function $\mathbf{u}$, defined by the metric parameters and target points (4.75). The l^∞ and l^1 errors between $\mathbf{u}$ and one of its approximations $\mathbf{u}_h^\varepsilon$ are respectively defined as

$$\max_{x \in \Omega_h} |\mathbf{u}_h^\varepsilon(x) - \mathbf{u}(x)|, \qquad h^2 \sum_{x \in \Omega_h} |\mathbf{u}_h^\varepsilon(x) - \mathbf{u}(x)|.$$

We display on Figure 4.3 the convergence curves for the centered $\mathcal{L}_h^\varepsilon$ (4.32) and the (unconditionally stable but less accurate) upwind scheme $\mathcal{L}_h^{\varepsilon,+}$ (4.38), and for $\varepsilon = \frac{1}{2}h^\alpha$ where $\alpha \in \{1/2, 2/3\}$. Empirically, the centered scheme works best when $\alpha = 2/3$, and the upwind scheme when $\alpha = 1/2$. This experiment illustrates and empirically confirms Corollary 4.3.14, which establishes that the minimal consistency error with the eikonal equation is achieved when $\varepsilon \approx h^\alpha$, where $\alpha = 2/3$ for the centered scheme, and $\alpha = 1/2$ for the upwind scheme. Note however that the empirical solution error appears to be higher than the scheme consistency error, which is $\mathcal{O}(h^\alpha)$, see Corollary 4.3.14.

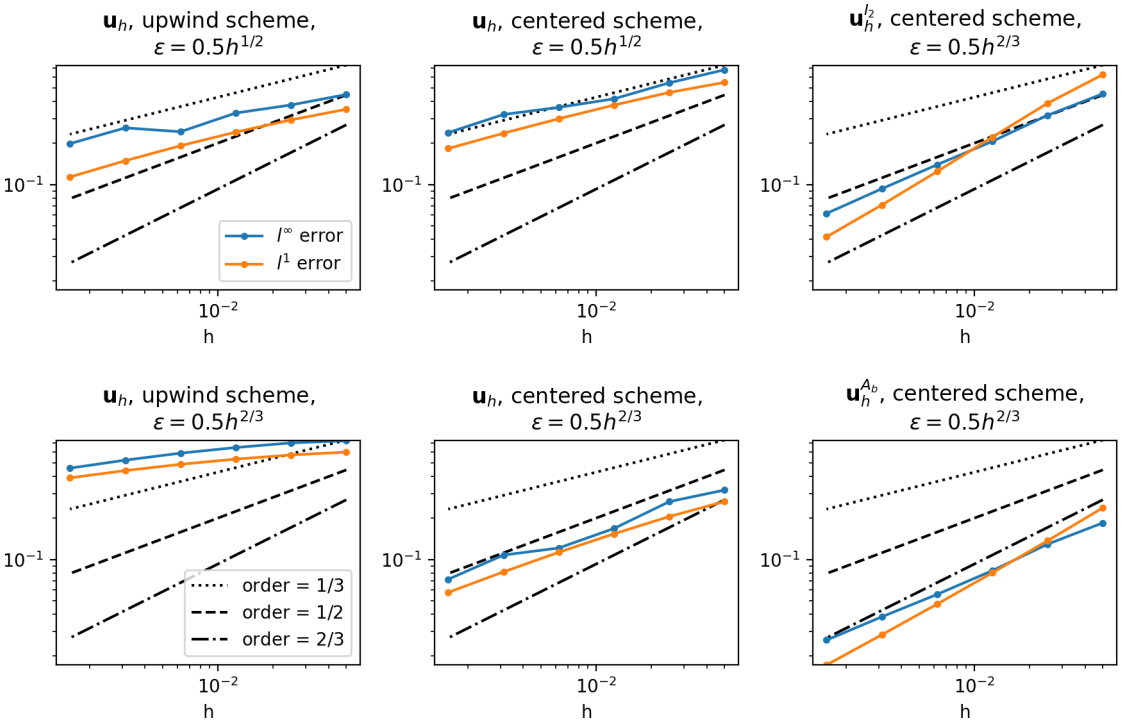


Figure 4.3: l^1 and l^∞ error between the exact distance $\mathbf{u}$, with parameters (4.75), and its numerical approximation, as a function of the grid scale h. Left : the upwind scheme $\mathcal{L}_h^{\varepsilon,+}$ (4.38) works best with $\varepsilon \approx h^{1/2}$. Center : the centered scheme is more accurate and works best with $\varepsilon \approx h^{2/3}$. The accuracy of the centered scheme solution is improved with a post-processing step, see Remark 4.3.2, which works best using the same stencil as the finite difference scheme (right, bottom), rather than an axis aligned stencil (right, top).

The post-processing step discussed in Remark 4.3.2, and adapted from [CWW13], allows to improve the accuracy of our numerical solution of the Randers eikonal equation solution, as illustrated on Figure 4.3 and 4.4. This post-processing works best when using the stencil of the finite difference scheme, as opposed to a basic axis-aligned stencil, see Figure 4.4 and the last sentence of Remark 4.3.2.

Optimal transport problems On Figure 4.5, we solve numerically the optimal transport problem (4.66), where μ and ν are uniform probability measures on $[-0.7, -0.1] \times [-0.5, 0.1]$ and $[0.1, 0.7] \times [-0.1, 0.5]$ respectively. We use Sinkhorn's algorithm (4.72) to numerically approximate the exponential Kantorovich potentials $\Phi, \Psi \in \mathbb{R}_+^{\Omega_h}$ maximizing (4.71), using the efficient approximation (4.73) of the product with the kernel $K_\varepsilon = \exp(-\operatorname{dist}_{\mathcal{F}}(x,y)/\varepsilon)$. The arrows on the figure follow Randers geodesics and illustrate a numerical approximation of the mapping $\sigma : \Omega_h \to \Omega$ defined by

$$\sigma(x) := \frac{1}{\mu_x} \sum_{y \in \Omega_h} P_{xy} y, \tag{4.78}$$

where $(P_{xy})_{x,y\in\Omega_h}$ is the optimal coupling measure (4.70) for the optimal transport problem (4.68). Thus $\sigma(x)$ is the barycenter of the image by the transport plan of the Dirac mass at x.

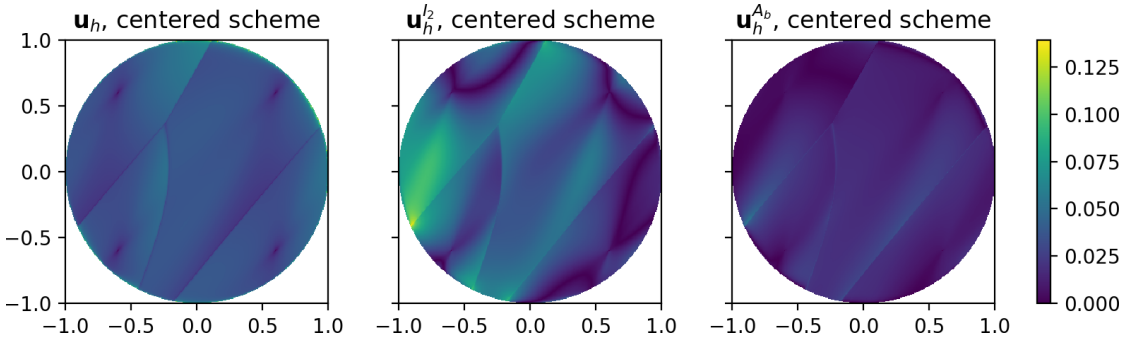


Figure 4.4: Absolute difference between the exact distance map $\mathbf{u}$ associated with the parameters (4.75) and its numerical approximation $\mathbf{u}_h^\varepsilon$ (left), the improved reconstruction using an axis-aligned stencil (center), or using the stencil of the finite difference scheme (right), see Remark 4.3.2. Grid scale $h = 0.0015625$ and $\varepsilon = 0.5h^{2/3}$.

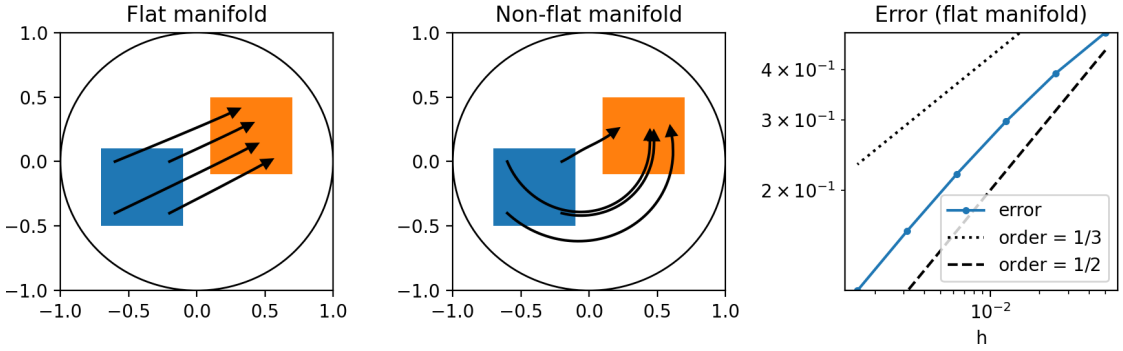


Figure 4.5: Numerical solution of the optimal transport problem (4.66). Left : manifold parameters (4.75), grid scale $h = 0.00625$. Middle : parameters (4.77), grid scale $h = 0.00625$. Right : convergence toward the exact Wasserstein distance as $h \to 0$, with parameters (4.75). In all cases: $\varepsilon = 0.5h^{\frac{2}{3}}$.

The numerical evaluation of σ involves a product with the kernel K_ε which again is efficiently approximated using (4.73). Note that the coupling measure P is typically not supported on a graph, not even approximately, and that σ is not a one to one mapping. In particular, σ does not approximate a translation in Figure (4.5, right). This behavior reflects the specific properties of the 1-Wasserstein distance, as opposed to the p-Wasserstein distance for $p > 1$, and it is not related to our numerical approximation procedure. Figure (4.5, right) displays the error between the approximation $W_h^\varepsilon(\mu, \nu)$ of the Wasserstein distance obtained with grid scale $h > 0$ and entropic relaxation $\varepsilon = \frac{1}{2}h^{\frac{2}{3}}$, and the exact optimal transport cost corresponding to the continuous problem without relaxation $\varepsilon = h = 0$.

4.7 Conclusions

In this paper, we introduced and studied a numerical scheme for approximating geodesic distances by solving a linear finite differences scheme, with an application to Schrödinger's entropic relaxation of the optimal transport problem. The approach builds on previous works [Var67;

CWW13; Cut13; Sol+15; YC16; Yan+18], and brings the following contributions: (i) justification of the distance computation method in the case of point sources, which is a common setting in applications, (ii) identification of the optimal parameter scaling $\varepsilon = h^{\frac{2}{3}}$, in contrast with the commonly used scaling $h = c\varepsilon$ which is inconsistent asymptotically (4.45), (iii) extension of these methods to asymmetric geometries defined by Randers metrics.

Our numerical scheme obeys the discrete degenerate ellipticity property, and thus benefits from comparison principles, numerical stability, and a convergence proof in the setting of viscosity solutions. For that purpose we use adaptive finite differences offsets depending on the PDE parameters and obtained via a tool from discrete geometry known as Selling's decomposition of positive definite matrices [Sel74; CS92]. Our convergence proof (in the case of a point source) exploits fine properties of Selling's decomposition: uniqueness, Lipschitz regularity, and spanning property (which implies the local connectivity of the stencils derived from it), for the first time in the context of PDE analysis [FM14; Mir18; Mir19]. Future work will be devoted to investigating their relevance in other applications to numerical analysis, and possible substitutes in dimension $d \geq 4$ where Selling's decomposition does not apply.

4.A Viscosity solutions

In this appendix, we establish the existence, uniqueness, comparison principles and convergence properties announced in section 4.2 for the following three PDEs:

$$u + 2\varepsilon\langle\nabla u, b\rangle - \varepsilon^2 \operatorname{Tr}(A_b\nabla^2 u) = 0 \text{ in } \Omega, \qquad u - \exp(-g/\varepsilon) = 0 \text{ on } \partial\Omega, \tag{4.79}$$

$$|\nabla \mathbf{u}|_A + \langle\nabla \mathbf{u}, b\rangle - 1 = 0 \text{ in } \Omega, \qquad \mathbf{u} - g = 0 \text{ on } \partial\Omega, \tag{4.80}$$

$$|\nabla \mathbf{u}|^2_{A_b} + 2\langle\nabla \mathbf{u}, b\rangle - \varepsilon \operatorname{Tr}(A_b\nabla^2 \mathbf{u}) - 1 = 0 \text{ in } \Omega, \qquad \mathbf{u} - g = 0 \text{ on } \partial\Omega, \tag{4.81}$$

The linear PDE (4.79), introduced in (4.3), is the foundation of our approach to Randers distance computation. The Randers eikonal PDE (4.80), which can be rephrased in many equivalent forms, see (4.5) and Corollary 4.2.9, characterizes Randers distance from the domain boundary with initial time penalty g. Finally (4.81) makes the link between the first two equations, being equivalent for any $\varepsilon > 0$ to (4.79) up to a logarithmic transformation of the unknown, and being equivalent for $\varepsilon = 0$ to (4.80). We recall that, by assumption, Ω is a bounded, connected and open domain with a $W^{3,\infty}$ boundary and $g \in C(\partial\Omega)$. The fields $A : \overline{\Omega} \to S_d^{++}$ and $b : \overline{\Omega} \to \mathbb{R}^d$ are Lipschitz, and $A_b := A - bb^\top$ is pointwise positive definite over $\overline{\Omega}$. The content of this section is presented in the appendix because it often mirrors similar results presented in the discrete setting of section 4.3 which we have chosen to emphasize, and because several key results are obtained by specialization of [BP88; BR98; BC97; CIL92]. We present in section 4.A.1 the concepts of degenerate elliptic operator and of viscosity solution to a PDE, and we justify the change of unknown known as the logarithmic transformation. The comparison principle, established in section 4.A.2 for the PDEs of interest, implies the uniqueness and boundedness of their solutions in Ω. We prove in section 4.A.3 the validity of the explicit solutions to (4.79) and (4.80) defined as a distance map (4.14) and as the expectation (4.23) of the stochastic process (4.22), and we establish convergence as $\varepsilon \to 0$.

4.A.1 Degenerate ellipticity, change of unknowns

The PDEs considered in this appendix (4.79) to (4.81) benefit from a common structure, known as degenerate ellipticity [CIL92; Obe06], introduced in Definition 4.A.1 below and whose discrete counterpart is presented in Definition 4.3.3.

Definition 4.A.1 (Degenerate ellipticity)**.** An operator $F\colon \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}_d \to \mathbb{R}$, denoted $F(x,t,p,X)$, is said *degenerate elliptic*[4] if it is (i) non-decreasing w.r.t. the second variable t, and (ii) non-increasing w.r.t. the last variable X for the Loewner order. The operator F is said elliptic if $F(x,t,p,X) - \delta t$ is degenerate elliptic for some constant $\delta > 0$.

The Dirichlet problem for a degenerate elliptic equation writes as

$$F(x, u(x), \nabla u(x), \nabla^2 u(x)) = 0 \text{ in } \Omega, \qquad u(x) - \psi(x) = 0 \text{ on } \partial\Omega, \tag{4.82}$$

where $\psi\colon \partial\Omega \to \mathbb{R}$. For example when considering equation (4.80), one should choose

$$F(x,t,p,X) = |p|_{A(x)} + \langle p, b(x)\rangle - 1, \qquad \psi(x) = g(x).$$

This specific operator F is degenerate elliptic, since $F(x,t,p,X)$ does not depend on either t or X, and thus obeys the required monotony conditions. Equation (4.81) is likewise defined by a degenerate elliptic operator, because the matrix field A_b is positive semi-definite. Equation (4.79) is elliptic thanks to the additional zeroth order term.

In the discrete setting, a comparison principle can be directly derived from the definition of ellipticity, see Lemma 4.3.4, and the related notions of sub-solution and super-solution are straightforward. Some additional care is however needed in the continuous case, see Definition 4.A.2, Proposition 4.A.7 and Theorem 4.A.8 below. For any bounded function $u\colon \overline{\Omega} \to \mathbb{R}^d$, we denote respectively by $u^*\colon \overline{\Omega} \to \mathbb{R}$ and $u_*\colon \overline{\Omega} \to \mathbb{R}$ its upper semicontinuous and lower semicontinuous envelopes, defined by

$$u^*(x) := \limsup_{y \in \overline{\Omega},\, y \to x} u(y), \qquad u_*(x) := \liminf_{y \in \overline{\Omega},\, y \to x} u(y). \tag{4.83}$$

Definition 4.A.2 (Viscosity solution)**.** Let $F\colon \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}_d \to \mathbb{R}$ be a continuous degenerate elliptic operator and let $\psi \in C(\partial\Omega)$. A bounded function $u\colon \overline{\Omega} \to \mathbb{R}$ is a *viscosity sub-solution* to (4.82) if for any $\varphi \in C^2(\overline{\Omega})$ and any local maximum $x \in \overline{\Omega}$ of $u^* - \varphi$,

$$\begin{cases} F(x, u^*(x), \nabla\varphi(x), \nabla^2\varphi(x)) \le 0 & \text{if } x \in \Omega, \\ \min\{u^*(x) - \psi(x),\, F(x, u^*(x), \nabla\varphi(x), \nabla^2\varphi(x)\} \le 0 & \text{if } x \in \partial\Omega. \end{cases}$$

It is a *viscosity super-solution* if for any $\varphi \in C^2(\overline{\Omega})$ and any local minimum $x \in \overline{\Omega}$ of $u_* - \varphi$,

$$\begin{cases} F(x, u_*(x), \nabla\varphi(x), \nabla^2\varphi(x)) \ge 0 & \text{if } x \in \Omega, \\ \max\{u_*(x) - \psi(x),\, F(x, u_*(x), \nabla\varphi(x), \nabla^2\varphi(x)\} \ge 0 & \text{if } x \in \partial\Omega. \end{cases}$$

It is a *viscosity solution* if it is both a viscosity sub-solution and super-solution.

Definition 4.A.2 encompasses discontinuous solutions u, obeying the boundary conditions in a weak sense, which allows implementing outflow boundary conditions in the case of the eikonal equation (4.80) by using large enough boundary data g. A well-known property of viscosity solutions is their stability under monotone changes of variables.

Proposition 4.A.3. *Let $F\colon \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}_d \to \mathbb{R}$ be a continuous degenerate elliptic operator, let $\psi \in C(\partial\Omega)$, let I, $J \subset \mathbb{R}$ be open intervals, let $\eta\colon I \to J$ be a strictly increasing C^2-diffeomorphism,*

[4] Or *proper degenerate elliptic* in the wording of [CIL92]. For consistency with the discrete case Definition 4.3.3, and following [Obe06], we drop the 'proper' qualifier.

and let $v\colon \overline{\Omega} \to I$ *be bounded away from* ∂I*. Define the continuous degenerate elliptic operator* $G\colon \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}_d \to \mathbb{R}$ *and boundary condition* $\chi : \partial\Omega \to \mathbb{R}$ *by*

$$G(x,t,p,X) := F(x,\, \eta(t),\, \eta'(t)p,\, \eta''(t)p\otimes p + \eta'(t)X), \qquad \chi(x) := \eta^{-1}(\psi(x)).$$

Then $u := \eta \circ v$ *is a viscosity sub-solution (respectively super-solution) to* (4.82) *if and only if* v *is a viscosity sub-solution (respectively super-solution) to*

$$G(x, v(x), \nabla v(x), \nabla^2 v(x)) = 0 \text{ in } \Omega, \qquad v(x) - \chi(x) = 0 \text{ on } \partial\Omega. \tag{4.84}$$

Proof. We only show the result for sub-solutions, since the case of super-solutions is similar. We assume that v is a sub-solution to (4.84) and prove that u is a sub-solution to (4.82). The proof of the converse is the same, using that

$$F(x,t,p,X) = G(x, \eta^{-1}(t), (\eta^{-1})'(t)p, (\eta^{-1})''(t)p\otimes p + (\eta^{-1})'(t)X).$$

The assumption that v is bounded away from ∂I implies that v^* and v_* are valued in I, hence $u^* = (\eta\circ v)^* = \eta \circ v^*$ is valued in J and likewise for u_*, by continuity of η. Let $\varphi \in C^2(\overline{\Omega})$ and $x \in \overline{\Omega}$ be a local maximum of $u^* - \varphi$. Without loss of generality, we may assume that $\varphi(\overline{\Omega}) \subset J$. Let $\tilde{\varphi} := \eta^{-1}\circ\varphi$. Using that η is strictly increasing, and $\varphi = \eta\circ\tilde{\varphi}$, we deduce that x is a local maximum of $v^* - \tilde{\varphi}$. We conclude the proof by noticing that for all $x \in \Omega$

$$F(x, u^*(x), \nabla\varphi(x), \nabla^2\varphi(x)) = G(x, v^*(x), \nabla\tilde{\varphi}(x), \nabla^2\tilde{\varphi}(x)).$$

In addition, if $x \in \partial\Omega$, then $u^*(x) - \psi(x)$ and $v^*(x) - \eta^{-1}(\psi(x))$ have the same sign. □

Remark 4.A.4. Sign changes exchange the notions of sub-solution and super-solution. More precisely, $u = -v$ is a viscosity sub-solution (resp. super-solution) to (4.82) iff v is a viscosity super-solution (resp. sub-solution) to (4.84) with

$$G(x,t,p,X) := -F(x,-t,-p,-X), \qquad \chi(x) = -\psi(x).$$

Combining Proposition 4.A.3 and Remark 4.A.4 allows to address the decreasing change of unknown $u = \exp(-\mathbf{u}/\varepsilon)$ considered by Varadhan [Var67], see Lemma 4.2.11. Note the discrete counterpart Proposition 4.3.10 of this result.

Corollary 4.A.5. *Let* $\mathbf{u} : \overline{\Omega} \to \mathbb{R}$*, and let* $u := \exp(-\mathbf{u}/\varepsilon)$*. Then* $\mathbf{u}$ *is a sub-solution (resp. super-solution) to* (4.81) *iff* u *is a super-solution (resp. sub-solution) to* (4.79)*.*

Proof. The PDE (4.79) corresponds to (4.82) with the following operator and boundary conditions

$$F(x,t,p,X) = t + 2\varepsilon\langle p, b(x)\rangle - \varepsilon^2 \operatorname{Tr}(A_b(x)X), \qquad \psi(x) = \exp(-g(x)/\varepsilon).$$

Applying successively Proposition 4.A.3 with the increasing diffeomorphism $\eta(t) := -\exp(-t/\varepsilon)$, and Remark 4.A.4, yields the boundary conditions $\chi(x) = -\eta^{-1}(\psi(x)) = g(x)$ and the operator

$$\begin{aligned} G(x,t,p,X) &= -F(x,-\eta(t),-\eta'(t)p,-\eta''(t)p\otimes p - \eta'(t)X) \\ &= -F\big(x, e^{-\frac{t}{\varepsilon}},\ -\tfrac{1}{\varepsilon}e^{-\frac{t}{\varepsilon}}p,\ \tfrac{1}{\varepsilon^2}e^{-\frac{t}{\varepsilon}}p\otimes p - \tfrac{1}{\varepsilon}e^{-\frac{t}{\varepsilon}}X\big) \\ &= -e^{-\frac{t}{\varepsilon}}\big(1 - 2\langle p, b(x)\rangle - \langle p, A_b(x)p\rangle + \varepsilon\operatorname{Tr}(A_b(x)X)\big). \end{aligned}$$

Simplifying by the positive factor $e^{-\frac{t}{\varepsilon}}$, and distributing the minus sign, we recognize (4.81). □

4.A.2 The comparison principle

The linear PDE (4.79) and Randers eikonal equation (4.80) admit a *strong comparison principle*, which in particular implies that their viscosity solutions are uniquely determined on Ω — though not on $\partial\Omega$. The proofs, presented in Proposition 4.A.7 and Theorem 4.A.8 below, are obtained as a specialization of [BR98]. For that purpose, we reformulate the first order term of (4.80) in Bellman form, based on the following identity: for all $x \in \overline{\Omega}$ and all $w \in \mathbb{R}^d$

$$|w|_{A(x)} + \langle w, b(x)\rangle = \sup_{\alpha\in\mathbb{B}^d} -\langle w, \mathfrak{b}(x,\alpha)\rangle, \qquad \text{where} \quad \mathfrak{b}(x,\alpha) := A^{\frac{1}{2}}(x)\alpha - b(x), \tag{4.85}$$

where $\mathbb{B}^d := \{x \in \mathbb{R}^d;\, \|x\| \leq 1\}$ denotes the closed unit ball.

Lemma 4.A.6. *The mappings $A^{\frac{1}{2}}, A_b^{\frac{1}{2}} : \overline{\Omega} \to S_d^{++}$ are Lipschitz continuous. The mapping $\mathfrak{b} : \overline{\Omega} \times \mathbb{B}^d \to \mathbb{R}^d$ defined by (4.85, right) is Lipschitz continuous. In addition, for each $x \in \overline{\Omega}$ and $p \in \mathbb{R}^d \setminus \{0\}$ there exists $\alpha \in \mathbb{B}^d$ such that $\langle \mathfrak{b}(x,\alpha), p\rangle > 0$.*

Proof. Recall that the mappings $A, A_b : \overline{\Omega} \mapsto S_d^{++}$ are Lipschitz continuous, and note that their lower eigenvalues are bounded away from zero by compactness. Since the matrix square root $\sqrt{\cdot} : S_d^{++} \to S_d^{++}$ is C^∞, as follows from holomorphic functional calculus[5], we obtain that $A^{1/2}$ and $A_b^{1/2}$ also are Lipschitz continuous on $\overline{\Omega}$. The announced regularity of $\mathfrak{b}$ follows.

Regarding the last property, we observe that choosing $\alpha = A^{\frac{1}{2}}(x)p/|A^{\frac{1}{2}}(x)p|$ yields

$$\langle \mathfrak{b}(x,\alpha),\, p\rangle = \langle \alpha - A^{-\frac{1}{2}}(x)b(x),\, A^{\frac{1}{2}}(x)p\rangle \geq \big(1 - |A^{-\frac{1}{2}}(x)b(x)|\big)|A^{\frac{1}{2}}(x)p| > 0, \tag{4.86}$$

since $|A^{-\frac{1}{2}}(x)b(x)| = |b(x)|_{A(x)^{-1}} < 1$ over $\overline{\Omega}$ by assumption. □

The comparison principle established in [BR98, Theorem 2.1] encompasses both the second order linear PDE (4.79), and the first order non-linear PDE (4.80) considered in this paper, although a reformulation is needed in the latter case.

Proposition 4.A.7. *Let $\overline{u}$ and $\underline{u}$ be respectively a sub-solution and a super-solution of the linear PDE* (4.79)*, for some $\varepsilon > 0$. Then $\overline{u}^* \leq \underline{u}_*$ in Ω.*

Proof. The announced result is a direct application of [BR98, Theorem 2.1], using that $A_b^{1/2} \colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ and $b \colon \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz continuous, $\partial\Omega$ is of class $W^{3,\infty}$, and $g \in C(\partial\Omega)$. □

Theorem 4.A.8. *Let $\overline{\mathbf{u}}$, $\underline{\mathbf{u}} : \overline{\Omega} \to \mathbb{R}$ be respectively a sub-solution and a super-solution of* (4.80)*. Then $\overline{\mathbf{u}}^* \leq \underline{\mathbf{u}}_*$ in Ω.*

Proof. Since (4.80) involves an operator which is degenerate elliptic but not elliptic, see Definition 4.A.1, we perform the Kruzhkov exponential change of variables and define $\overline{\mathbf{v}} := -\exp(-\overline{\mathbf{u}})$ and $\underline{\mathbf{v}} := -\exp(-\underline{\mathbf{u}})$. By Proposition 4.A.3, $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$ are respectively a viscosity sub-solution and super-solution to

$$|\nabla \mathbf{v}(x)|_{A(x)} + \langle \nabla \mathbf{v}(x), b(x)\rangle + \mathbf{v}(x) = 0 \text{ in } \Omega, \qquad \mathbf{v}(x) + \exp(-g(x)) = 0 \text{ on } \partial\Omega.$$

The boundary $\partial\Omega$ is of class $W^{3,\infty}$, and the boundary data $-\exp(-g) \in C(\partial\Omega)$, consistently with the framework of [BR98]. Furthermore, the PDE can be rewritten as $\sup_{\alpha\in\mathbb{B}^d} -\langle \mathfrak{b}(x,\alpha), \nabla \mathbf{v}(x)\rangle + \mathbf{v}(x) = 0$ in Ω, and the required regularity properties of $\mathfrak{b}$ are established in Lemma 4.A.6, as well as the additional condition which amounts to a local controllability property. Then by [BR98, Theorem 2.1], we obtain $\overline{\mathbf{v}}^* \leq \underline{\mathbf{v}}_*$ in Ω, and therefore $\overline{\mathbf{u}}^* \leq \underline{\mathbf{u}}_*$ in Ω as announced. □

[5] More directly, if the eigenvalues of $A \in S_d^{++}$ lie in $]0, 2r[$, then one has the series expansion $\sqrt{A} = \sqrt{r}\sum_{k\geq 0} a_k (A/r - \mathrm{Id})^k$, where $\sqrt{1+t} = \sum_{k\geq 0} a_k t^k$ for all $t \in]-1, 1[$.

4.A.3 Explicit solutions, and convergence

We establish that viscosity solutions to Randers eikonal equation (4.80) and to the linear PDE (4.79) may be explicitly obtained as the distance from the boundary (4.4) with suitable penalty term, and as the expectation of a stochastic process (4.23). We also prove bounds for these solutions, see Theorems 4.A.9 and 4.A.11, and conclude the proof of Varadhan's formula for Randers metrics in Theorem 4.A.12.

Theorem 4.A.9. *Denote by $\mathcal{F}$ the Randers metric of parameters (M, ω) dual to (A, b), see Lemma 4.2.6. Then $\mathbf{u} : x \in \overline{\Omega} \mapsto \min_{p\in\partial\Omega} \operatorname{dist}_{\mathcal{F}}(p, x) + g(p)$ is a bounded viscosity solution to (4.80).*

Proof. The boundedness of $\mathbf{u}$ follows from the equivalence of the Randers distance with the Euclidean distance, see Lemma 4.2.4. Since $g \in C(\partial\Omega)$ and the control function $\mathfrak{b}$ is Lipschitz continuous [BC97, Theorem V.4.13 and Remark V.4.14] yields a viscosity solution $\mathbf{v}$ to (4.80) in the form

$$\mathbf{v}(x) = \inf\{T + g(\gamma_x^\alpha(T));\, T \geq 0,\, \alpha : [0, T] \to \mathbb{B}^d, \gamma_x^\alpha(T) \in \partial\Omega\} \tag{4.87}$$

where $\gamma = \gamma_x^\alpha$ is defined by $\gamma(0) = x$ and $\gamma'(t) = \mathfrak{b}(\gamma(t), \alpha(t))$ for all $0 \leq t \leq T$, and where α is implicitly assumed to be measurable. Now, for any $v \in \mathbb{R}^d$ one obtains, omitting the argument x in $M(x)$, $\omega(x)$, $A(x)$ and $b(x)$ for readability

$$\begin{aligned}\mathcal{F}_x(v) \leq 1 &\Leftrightarrow |v|_M + \langle \omega, v\rangle \leq 1 \Leftrightarrow |v - b|_{A^{-1}} \leq 1 \Leftrightarrow \exists \tilde{\alpha} \in \mathbb{B}^d, v - b = A^{\frac{1}{2}}\tilde{\alpha} \\ &\Leftrightarrow \exists \alpha \in \mathbb{B}^d, v = -\mathfrak{b}(x, \alpha),\end{aligned}$$

where the first equivalence holds by definition, the second is established in Lemma 4.2.7, the third follows from $|A^{\frac{1}{2}}\tilde{\alpha}|_{A^{-1}} = |\tilde{\alpha}|$ for any $\tilde{\alpha} \in \mathbb{R}^d$, and the last is obtained by choosing $\alpha = -\tilde{\alpha}$. Thus

$$\begin{aligned}\mathbf{v}(x) = \inf\{T + g(\gamma(T)); T \geq 0, \exists \gamma \in \operatorname{Lip}([0, T], \overline{\Omega}), \gamma(0) = x, \gamma(T) \in \partial\Omega, \\ \mathcal{F}_{\gamma(t)}(-\gamma'(t)) \leq 1, \text{ for a.e. } t \in [0, T]\}.\end{aligned}$$

Noting that any Lipschitz path can be reparametrized at constant speed w.r.t. the metric $\mathcal{F}$, and have its orientation reversed (from x to $\partial\Omega$), we obtain that $\mathbf{v}(x) = \mathbf{u}(x)$, which concludes the proof. □

We obtain a sub-solution and a super-solution to the PDE (4.81), independent of the relaxation parameter, similarly to the discrete case in Lemma 4.3.15

Lemma 4.A.10. *The PDE (4.81) admits, for any $\varepsilon \geq 0$, the constant sub-solution $\overline{\mathbf{u}} : x \in \overline{\Omega} \mapsto g_{\min}$, where $g_{\min} := \min\{g(y);\, y \in \partial\Omega\}$. It also admits the affine super-solution $\underline{\mathbf{u}} : x \in \overline{\Omega} \mapsto \langle p, x\rangle + c_{\max}$, for any $p \in \mathbb{R}^d$ such that $|p|$ is sufficiently large, where $c_{\max} := \max\{g(y) - \langle p, y\rangle;\, y \in \partial\Omega\}$.*

Proof. Denote $\mathcal{S}^\varepsilon\mathbf{u} := |\nabla\mathbf{u}|^2_{A_b} + 2\langle\nabla\mathbf{u}, b\rangle - \varepsilon \operatorname{Tr}(A_b\nabla^2\mathbf{u}) - 1$ the operator of (4.81). Clearly $\mathcal{S}^\varepsilon\overline{\mathbf{u}} = -1 < 0$ in $\overline{\Omega}$, whereas $\mathcal{S}^\varepsilon\underline{\mathbf{u}}(x) = |p|^2_{A_b(x)} + 2\langle p, b(x)\rangle - 1 \geq c_0 > 0$ for all $x \in \overline{\Omega}$, provided $|p|$ is sufficiently large, since A_b and b are bounded over $\overline{\Omega}$, and A_b is uniformly positive definite. The constants $g_{\min}$ and $c_{\max}$ are chosen so as to comply with the boundary conditions. □

Theorem 4.A.11. *For any $\varepsilon > 0$, the function $u_\varepsilon : \overline{\Omega} \to \mathbb{R}_-$ defined by (4.23) is a viscosity solution to (4.79). In addition, u_ε is positive, and $\overline{\mathbf{u}} \leq \mathbf{u}_\varepsilon \leq \underline{\mathbf{u}}$ in Ω, where $\mathbf{u}_\varepsilon := -\varepsilon \ln(u_\varepsilon)$ and $\overline{\mathbf{u}}$ and $\underline{\mathbf{u}}$ are from Lemma 4.A.10.*

Proof. Since $A_b^{1/2}\colon \mathbb{R}^d \to \mathcal{S}_d^{++}$ and $b\colon \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz continuous, $\partial\Omega$ is of class $W^{3,\infty}$, and $g \in C(\partial\Omega)$, [BR98, Theorem 3.1] implies that u_ε is a viscosity solution to (4.79).

By Corollary 4.A.5, $\overline{u}_\varepsilon := \exp(-\underline{\mathbf{u}}/\varepsilon)$ and $\underline{u}_\varepsilon := \exp(-\overline{\mathbf{u}}/\varepsilon)$ are respectively a sub-solution and a super-solution to (4.79). Thus $\overline{u}_\varepsilon \le (u_\varepsilon)_* \le u_\varepsilon \le (u_\varepsilon)^* \le \underline{u}_\varepsilon$ in Ω by Theorem 4.A.8. Therefore u_ε is positive, as announced, and we conclude using the monotony of the logarithm. □

We are able to complete the proof of formula (4.24) by making rigorous the passing to the limit between problems (4.81) and (4.80). Note that we follow a standard sketch of proof, already used in [BP88, Proposition II.6] for example.

Theorem 4.A.12. *With the notations of Theorem 4.A.11, and denoting by* $\mathbf{u}$ *the solution to* (4.15)*, one has* $\mathbf{u}_\varepsilon \to \mathbf{u}$ *uniformly on compact subsets of* Ω*, as* $\varepsilon \to 0$*.*

Proof. By Theorem 4.A.11, $\mathbf{u}_\varepsilon$ is bounded above and below, uniformly on $\overline{\Omega}$ and uniformly w.r.t. $\varepsilon > 0$. Therefore the following limit is well-defined, for any $x \in \overline{\Omega}$

$$\overline{\mathbf{v}}(x) := \limsup_{\varepsilon\to 0, y\to x} \mathbf{u}_\varepsilon(y) \quad \Big(= \lim_{\delta\to 0} \sup\big\{\mathbf{u}_\eta(y);\, 0<\eta\le\delta,\, |y-x|\le\delta\big\}\Big),$$

and likewise $\underline{\mathbf{v}}(x) := \liminf \mathbf{u}_\varepsilon(y)$ as $\varepsilon \to 0$ and $y \to x$. Thus we can apply [CIL92, Lemma 6.1 and Remark 6.3] to functions $(\mathbf{u}_\varepsilon)_*$ and $(\mathbf{u}_\varepsilon)^*$, and deduce that $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$ are respectively a viscosity subsolution and supersolution to (4.81) with $\varepsilon = 0$, or equivalently to (4.80) by Corollary 4.2.9. Hence by Theorem 4.A.8, $\overline{\mathbf{v}} \le \mathbf{u}_* \le \mathbf{u}^* \le \underline{\mathbf{v}}$ on Ω. By definition, $\overline{\mathbf{v}} \ge \underline{\mathbf{v}}$ on $\overline{\Omega}$. Therefore $\overline{\mathbf{v}} = \mathbf{v} = \underline{\mathbf{v}}$ on Ω. The locally uniform convergence of $\mathbf{u}_\varepsilon$ to $\mathbf{u}$ on Ω follows from the definitions of $\overline{\mathbf{v}}$ and $\underline{\mathbf{v}}$. □

4.B Selling's decomposition of positive definite matrices

This appendix is devoted to a brief description of Selling's decomposition of symmetric positive definite matrices [Sel74; CS92] of dimension $d \in \{2, 3\}$, a tool from algorithmic geometry which is convenient when discretizing anisotropic PDEs on Cartesian grids ([FM14; Mir18; Mir19] and Chapter 2), here used section 4.3.1. Selling's formula and algorithm are presented in Lemma 4.B.2 and Proposition 4.B.3. Two properties of the resulting normalized decomposition (4.91), established in Propositions 4.B.5 and 4.B.8, are used in section 4.4.3 for the first time in the context of PDE numerical analysis.

Definition 4.B.1. A superbase of $\mathbb{Z}^d$ is a family $(v_0, \cdots, v_d) \in (\mathbb{Z}^d)^{d+1}$ such that $v_0 + \cdots + v_d = 0$ and $|\det(v_1, \cdots, v_d)| = 1$. It is said D-obtuse, where $D \in S_d^{++}$, iff $\langle v_i, Dv_j\rangle \le 0$ for all $0 \le i < j \le d$.

To each superbase $(v_0, \cdots, v_d)$ of $\mathbb{Z}^d$, we associate the family of vectors $e_{ij} \in \mathbb{Z}^d$, $0 \le i < j \le d$ defined by the linear relations

$$\langle e_{ij}, v_k\rangle = \delta_{ik} - \delta_{jk}, \tag{4.88}$$

for all $0 \le k \le d$, where δ_{ij} denotes Kronecker's symbol. In dimension $d = 2$ (resp. $d = 3$), if $\{i, j, k\} = \{0, 1, 2\}$ (resp. $\{i, j, k, l\} = \{0, 1, 2, 3\}$), one easily checks that $e_{ij} = \pm v_k^\perp$ (resp. $e_{ij} = \pm v_k \times v_l$). Selling's formula and algorithm are classical [Sel74; CS92; Mir18], yet their (short) proofs are presented for completeness, since they are core elements of our numerical scheme.

Lemma 4.B.2 (Selling's formula)**.** *Let* $D \in S_d$ *and let* $(v_0, \cdots, v_d)$ *be a superbase of* $\mathbb{Z}^d$*. Then*

$$D = -\sum_{0\le i<j\le d} \langle v_i, Dv_j\rangle e_{ij} e_{ij}^\top. \tag{4.89}$$

Proof. By (4.88) we obtain $\langle v_i, Dv_j\rangle = \langle v_i, D'v_j\rangle$ for all $0 \le i < j \le d$, where D' denotes (4.89, rhs). Thus $\langle v_i, Dv_i\rangle = \langle v_i, D'v_i\rangle$ by linearity and since $v_i = -(v_0 + \cdots + v_{i-1} + v_{i+1} + \cdots + v_d)$. The result follows since $(v_1, \cdots, v_d)$ is a basis. □

If $D \in S_d^{++}$ and $(v_0, \cdots, v_d)$ is D-obtuse, then (4.89) is known as Selling's decomposition of D. Selling's algorithm provides a constructive proof of existence of such a D-obtuse superbase, in dimension $d \in \{2, 3\}$.

Proposition 4.B.3 (Selling algorithm)**.** *Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$, $d \in \{2, 3\}$, and let $D \in S_d^{++}$. If b is not D-obtuse, permute it so that $\langle v_0, Dv_1\rangle > 0$ and update it as follows*

$$b \leftarrow (-v_0, v_1, v_0 - v_1) \text{ if } d = 2, \qquad b \leftarrow (-v_0, v_1, v_2 + v_0, v_3 + v_0) \text{ if } d = 3. \tag{4.90}$$

Repeating this operation yields a D-obtuse superbase in finitely many steps.

Proof. Define $\mathcal{E}(b) = \sum_{i=0}^d \|v_i\|_D^2$. If $b = (v_0, \cdots, v_d)$ is such that $\delta := \langle v_0, Dv_1\rangle > 0$, and if b' is defined by (4.90) then one easily checks that b' also is a superbase and that $\mathcal{E}(b') = \mathcal{E}(b) - C_d\delta$, where $C_2 = 4$ and $C_3 = 2$. There are only finitely many superbases of $\mathbb{Z}^d$ whose energy $\mathcal{E}$ is below any given bound, since their elements have integer coordinates and since D is positive definite. Hence Selling's algorithm must terminate, which happens when the iteration condition fails, i.e. when a D-obtuse superbase b is obtained. This concludes the proof. □

The elements of a D-obtuse superbase, and the corresponding offsets in Selling's formula, are bounded in terms of the anisotropy ratio $\mu(D) := \sqrt{\|D\|\|D^{-1}\|}$.

Proposition 4.B.4. *Let $D \in S_d^{++}$, and let $b = (v_0, \cdots, v_d)$ be a D-obtuse superbase, where $d \in \{2, 3\}$. Then $|v_i| \le C\mu(D)$, $0 < i < d$, and $|e_{ij}| \le 2C\mu(D)$, $0 \le i < j \le d$, where $C = 2$ if $d = 2$ (resp. $C = 2\sqrt{3}$ if $d = 3$). In fact, one has the slightly stronger estimates $|v_i|_D \le C\|D\|^{\frac{1}{2}}$ and $|e_{ij}|_{D^{-1}} \le 2C\|D^{-1}\|^{\frac{1}{2}}$.*

Proof. The bounds $|v_i| \le C\mu(D)$ and $|e_{ij}| \le 2C\mu(D)$ are established in [Mir18, Proposition 4.8 and Theorem 4.11]. Inspecting the proof of these results, one obtains the other announced estimates. Specifically, $|v_i|_D \le C\|D\|^{\frac{1}{2}}$ is established in the last line of [Mir18, Proposition 4.8]. Using this refined estimate (instead of $|v_i| \le C\mu(D)$) in the proof of [Mir18, Theorem 4.11] yields $|e_{ij}|_{D^{-1}} \le 2C\|D^{-1}\|^{\frac{1}{2}}$ (instead of $|e_{ij}| \le 2C\mu(D)$). The result follows. □

Selling's decomposition of a matrix $D \in S_d^{++}$, $d \in \{2, 3\}$, is obtained by applying Selling's formula Lemma 4.B.2 to a D-obtuse superbase, whose existence is ensured by Selling's algorithm Proposition 4.B.3. This description is constructive and used in all our numerical experiments, since it is efficient enough for the moderately ill-conditioned matrices encountered in our applications. We normalize Selling's decomposition as follows, up to replacing some offsets with their opposites:

$$D = \sum_{e \in \mathcal{Z}^d} \rho(e; D)\, ee^\top, \qquad \text{where } \mathcal{Z}^d := \{e \in \mathbb{Z}^d;\ e \succ_{\text{lex}} 0\}, \tag{4.91}$$

where $\succ_{\text{lex}}$ stands for the lexicographic ordering. (Note that exactly one of $e \succ_{\text{lex}} 0$ or $-e \succ_{\text{lex}} 0$ holds for each $e \in \mathbb{Z}^d \setminus \{0\}$.) The weights $[\rho(e; D)]_{e \in \mathcal{Z}^d}$ are known as Selling parameters [CS92], and depend on D but *not* on the choice of D-obtuse superbase, see e.g. Remark 2.2.13 for a proof. In view of Selling's formula (4.89), one has $\rho(e; D) = 0$ except for at most $d(d+1)/2$ offsets $e \in \mathcal{Z}^d$. In addition, $\rho(e; D) = 0$ if $|e| > 2C\mu(D)$, by Proposition 4.B.4.

Proposition 4.B.5 (Lipschitz regularity)**.** *For any $e \in \mathcal{Z}^d$, $d \in \{2, 3\}$, the mapping $D \in S_d^{++} \mapsto \rho(e; D)$ is locally Lipschitz with constant $C^2\mu(D)^2$, where C is from Proposition 4.B.4.*

Proof. Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$, and define $S_b := \{D \in S_d^{++};\ b \text{ is } D\text{-obtuse}\}$. For each $0 \leq i < j \leq d$ let $\tilde{e}_{ij} := \pm e_{ij}$, where the sign is chosen so that $\tilde{e}_{ij} \in \mathcal{Z}^d$. By (4.89) one has $\rho(D; \tilde{e}_{ij}) = -\langle v_i, Dv_j\rangle$ for all $D \in S_b$, which is a linear function of D with Lipschitz constant at most $|v_i||v_j| \leq C^2\mu(D)^2$ by Proposition 4.B.4. In addition, $\rho(D; e) = 0$ for all $D \in S_b$ and all $e \in \mathcal{Z}^d \setminus \{\tilde{e}_{ij}; 0 \leq i < j \leq d\}$, thus $D \mapsto \rho(e; D)$ is Lipschitz with the announced constant over the set S_b. The announced result follows since S_d^{++} is the union of the closed and convex sets S_b associated to superbases b of $\mathbb{Z}^d$, by Proposition 4.B.3, and since this union is locally finite by Proposition 4.B.4 □

We conclude this appendix by establishing, in Proposition 4.B.8, that some offsets of Selling's decomposition, associated with weights suitably bounded below, span the integer lattice $\mathbb{Z}^d$ by linear combinations with integer coefficients. This implies that the stencils of our numerical scheme (4.32) define a locally connected graph, a property used in section 4.4.3 to control its solution in the neighborhood of a point source.

Lemma 4.B.6. *Let $(v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$, and let $(i_k, j_k)_{k=1}^d$ be such that $0 \leq i_k < j_k \leq d$ for all $0 \leq k \leq d$. Then $\det(e_{i_1 j_1}, \cdots, e_{i_d j_d}) \in \{-1, 0, 1\}$.*

Proof. By Definition 4.B.1, $(v_1, \cdots, v_d)$ is a basis of $\mathbb{Z}^d$. We may thus assume that $(v_1, \cdots, v_d)$ is the canonical basis of $\mathbb{Z}^d$, up to a change of basis, so that $v_0 = (-1, \cdots, -1)^\top$. Then $e_{0j} = -v_j$ for all $1 \leq j \leq d$, and $e_{ij} = v_i - v_j$ for all $1 \leq i < j \leq d$. Each of the vectors e_{ij}, $0 \leq i < j \leq d$, thus features at most once the coefficient 1, and at most once the coefficient -1, the other coefficients being 0. The announced result then follows from [BG15, Proposition 2.37]. □

Lemma 4.B.7. *Let $D \in S_d^{++}$, and let $e_1, \cdots, e_I \in \mathbb{R}^d$ be such that $D = \sum_{i=1}^I e_i e_i^\top$. Then there exists $1 \leq i_1 < \cdots < i_d \leq I$ s.t. $\sum_{k=1}^d e_{i_k} e_{i_k}^\top \geq cD$, where $c = c(d, I) > 0$.*

Proof. Without loss of generality, up to a linear change of coordinates, one may assume that $D = \mathrm{Id}$ is the $d \times d$ identity matrix. Define

$$\Xi := \big\{(e_i)_{i=1}^I \in (\mathbb{R}^d)^I;\ \sum_{1 \leq i \leq I} e_i e_i^\top = \mathrm{Id}\,\big\}, \qquad \Lambda\big((e_i)_{i=1}^I\big) = \max_{i_1 < \cdots < i_d} \lambda_{\min}\Big(\sum_{1 \leq k \leq d} e_{i_k} e_{i_k}^\top\Big),$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. Any family $(e_i)_{i=1}^I \in \Xi$ spans $\mathbb{R}^d$, thus a basis $(e_{i_1}, \cdots, e_{i_d})$ can be extracted from it, and therefore $\Lambda((e_i)_{i=1}^I) \geq \lambda_{\min}(\sum_{k=1}^d e_{i_k} e_{i_k}^\top) > 0$. Denoting by $c(I, d)$ the lower bound of Λ over Ξ, which is positive since Ξ is compact and since Λ is continuous and positive over Ξ, we conclude the proof. □

Proposition 4.B.8 (Spanning property)**.** *For any $D \in S_d^{++}$, $d \in \{2, 3\}$, there exists $e_1, \cdots, e_d \in \mathcal{Z}^d$ such that, for some absolute constant $c > 0$*

$$\det(e_1, \cdots, e_d) = 1, \qquad \min_{1 \leq i \leq d} \rho(e_i; D) \geq c\|D^{-1}\|^{-1}. \tag{4.92}$$

Proof. From (4.91) and Lemma 4.B.7 there exists $e_1, \cdots, e_d \in \mathcal{Z}^d$ such that $\sum_{i=1}^d \rho_i e_i e_i^\top \geq cD$, where $\rho_i := \rho(e_i; D)$ and $c = c(d, I) > 0$ is an absolute constant since $d \in \{2, 3\}$ and $I = d(d+1)/2$. Let v be a non-zero vector orthogonal to $e_2, \cdots, e_d$. Then $c|v|_D^2 \leq \rho_1 \langle v, e_1\rangle^2 \leq \rho_1 |v|_D^2 |e_1|_{D^{-1}}^2 \leq (2C)^2 \rho_1 |v|_D^2 \|D^{-1}\|$ by Proposition 4.B.4. Thus $\rho_1 \geq (c/(2C)^2)\|D^{-1}\|^{-1}$, and likewise for $\rho_2, \cdots, \rho_d$, which concludes the proof. □

Chapter 5

Monotone and second order consistent scheme for the two-dimensional Pucci equation

This chapter corresponds to the paper [BBM21b].

5.1 Introduction

Degenerate Ellipticity (DE) is a property of a class of partial differential operators, often non-linear and of order at most two. When satisfied, it implies a generalized comparison principle, from which can be deduced the existence, uniqueness and stability of a viscosity solution to the Partial Differential Equation (PDE), under mild additional assumptions [CIL92]. Discrete degenerate ellipticity is the corresponding property for numerical schemes, see Definition 5.2.3, which has similarly strong implications and often turns the convergence analysis of solutions into a simple verification [Obe06]. It is therefore appealing to design PDE discretizations preserving the DE property, yet a strong limitation of the current approaches [BS91; Obe08; FJ17] is their low consistency order, usually below one. Filtered schemes [FO13] attempt to mitigate this issue by combining a DE scheme of low consistency order with a non-DE scheme of high consistency order, but their use requires careful parameter tuning, and theoretical results are lacking regarding their effective accuracy.

In this paper, we propose a new approach to develop second order accurate DE schemes, which is the highest achievable consistency order [Obe06], on two-dimensional Cartesian grids. The operator must be given in Bellman form as follows

$$\Lambda u(x) = \sup_{\alpha \in \mathcal{A}} a_\alpha + b_\alpha u(x) - \mathrm{Tr}(D_\alpha \nabla^2 u(x)), \tag{5.1}$$

where $\mathcal{A}$ is an abstract set of parameters, and the coefficients $a_\alpha \in \mathbb{R}$, $b_\alpha \geq 0$, and symmetric positive definite matrix D_α may additionally depend on the position x. A specific feature of our approach, that is tied to the structure of the addressed problems, is that the parameter space $\mathcal{A}$ is not discretized. We apply this approach to the two-dimensional Pucci equation:

$$\lambda_{\min}(\nabla^2 u(x)) + \mu\, \lambda_{\max}(\nabla^2 u(x)) = f(x), \tag{5.2}$$

with Dirichlet boundary conditions, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of a symmetric matrix, and where $\mu > 0$. This PDE admits the following Bellman form, when $\mu \leq 1$, which we assume for simplicity:

$$\max_{\theta\in[0,\pi]} -\operatorname{Tr}(D(\theta,\mu)\nabla^2 u(x)) = -f(x), \qquad \text{where } D(\theta,\mu) := R_\theta \begin{pmatrix} 1 & 0 \\ 0 & \mu \end{pmatrix} R_\theta^T, \tag{5.3}$$

and where $R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ denotes the rotation matrix of angle $\theta \in \mathbb{R}$. Our approach also applies in the case $\mu \geq 1$, with only the slight modification that the max in (5.3) is replaced with a min. Note that the optimization space in (5.3) is $\mathcal{A} = [0, \pi]$, which is compact and one dimensional, thus easing the theoretical study and the numerical implementation.

Motivation for this study. The Pucci equation interpolates between two fundamental problems in analysis: the Poisson problem when $\mu = 1$, and the (lower-)convex envelope of the boundary conditions when $\mu = 0$ and $f = 0$. It is also an excellent representative of the class of Pucci extremal operators, a.k.a. operators which can be written in the form (5.1), perhaps replacing the inf with a sup. This class also encompasses the Monge-Ampère operator, known for its applications in optimal transport and optics, to which similar techniques may be applied [BCM16].

5.2 Discretization

We rely on a tool from algorithmic lattice geometry, known as Selling's formula §5.2.1, which is particularly adequate for discretizing degenerate elliptic PDEs on Cartesian grids of dimension two [BOZ04] or three [Mir18; Mir19; FM14]. Throughout this section $\Omega \subset \mathbb{R}^2$ denotes a bounded domain, and $h > 0$ a grid scale. Define

$$\Omega_h := h\mathbb{Z}^2 \cap \Omega, \qquad \Delta_h^e u(x) := \frac{u(x+he) - 2u(x) + u(x-he)}{h^2}, \tag{5.4}$$

the discrete domain and the second order finite difference of a map $u : \Omega_h \cup \partial\Omega \to \mathbb{R}$ at $x \in \Omega_h$ in the direction $e \in \mathbb{Z}^2$. When x is adjacent to $\partial\Omega$ the latter formula becomes

$$\Delta_h^e u(x) := \frac{2}{h_+ + h_-}\Big(\frac{u(x+h_+e) - u(x)}{h_+} + \frac{u(x-h_-e) - u(x)}{h_-}\Big), \tag{5.5}$$

where $h_\pm > 0$ is the least value such that $x \pm h_\pm e \in \Omega_h \cup \partial\Omega$. Note that (5.4, right) is a second order consistent approximation of $\langle e, \nabla^2 u(x) e\rangle$, whereas (5.5) is only first order consistent. Thus

$$\operatorname{Tr}(ee^T\nabla^2 u(x)) = \langle e, \nabla^2 u(x) e\rangle = \Delta_h^e u(x) + \mathcal{O}(h^r), \tag{5.6}$$

where $r = 1$ if x is adjacent to $\partial\Omega$, and $r = 2$ otherwise.

5.2.1 Selling's formula

Selling's decomposition of an element of the set S_2^{++} of symmetric positive definite 2×2 matrices, see Proposition 5.2.2, can be regarded as a variant of the eigenvector/eigenvalue decomposition, but whose vectors have *integer entries*. We rely on it to discretize non-divergence form linear (5.8) and non-linear (5.11) operators, in a manner that achieves discrete degenerate ellipticity, see Definition 5.2.3.

Definition 5.2.1. A superbase of $\mathbb{Z}^2$ is a triplet $(e_0, e_1, e_2) \in (\mathbb{Z}^2)^3$ such that $e_0 + e_1 + e_2 = 0$ and $|\det(e_1, e_2)| = 1$. It is D-obtuse, where $D \in S_2^{++}$, iff $\langle e_i, De_j\rangle \leq 0$ for all $i \neq j$.

Proposition 5.2.2 (Selling [Sel74]). *For each $D \in S_2^{++}$ there exists a D-obtuse superbase (e_0, e_1, e_2) of $\mathbb{Z}^2$, which can be obtained from Selling's algorithm. Furthermore one has Selling's formula*

$$D = \sum_{0 \le i \le 2} \rho_i v_i v_i^\top \qquad \text{with} \quad \rho_i := -\langle e_{i-1}, D e_{i+1} \rangle \ge 0, \quad v_i := e_i^\perp \in \mathbb{Z}^2, \tag{5.7}$$

where $e^\perp := (-b, a)^\top$ if $e = (a, b)^\top \in \mathbb{R}^2$. The set $\{(\rho_i, \pm v_i);\ 0 \le i \le 2,\ \rho_i > 0\}$ is uniquely determined. In (5.7), the indices $i-1$ and $i+1$ are understood modulo 3.

Based on this formula, one can consider the following finite differences operator:

$$\Delta_h^D u(x) := \sum_{0 \le i \le 2} \rho_i \Delta_h^{v_i} u(x). \tag{5.8}$$

Using (5.6), (5.7), (5.8) and the linearity of the trace operator on matrices, we obtain

$$\operatorname{Tr}(D \nabla^2 u(x)) = \sum_{0 \le i \le 2} \rho_i \operatorname{Tr}(v_i v_i^T \nabla^2 u(x)) = \Delta_D^h u(x) + \mathcal{O}(h^r),$$

where again $r = 1$ if x is adjacent to $\partial\Omega$, and $r = 2$ otherwise.

We illustrate on Figure 5.1 the relation between the anisotropy defined by a symmetric positive definite matrix $D \in S_2^{++}$, and the corresponding offsets $\pm v_0, \pm v_1, \pm v_2 \in \mathbb{Z}^2$ in Selling's formula. (The weights ρ_i are illustrated on Figure 5.2.) For that purpose, we rely on a parametrization $\mathbf{D}$ of the 2×2 symmetric positive definite matrices of unit trace, by the points (x, y) of the open unit ball:

$$\mathbf{D}(x, y) := \frac{1}{2} \begin{pmatrix} 1 + x & y \\ y & 1 - x \end{pmatrix}, \qquad \text{where } x^2 + y^2 < 1. \tag{5.9}$$

This parametrization is closely related to the Pauli matrices in quantum mechanics. A $\mathbf{D}(x, y)$-obtuse superbase is known explicitly, depending on a suitable triangulation of the unit disc, see Figure (5.1, right).

Definition 5.2.3 (Discrete degenerate ellipticity [Obe06]). A numerical scheme on a finite set X is a map $F : U \to U$, where $U := \mathbb{R}^X$ is the set of functions from X to $\mathbb{R}$, of the form:

$$Fu(x) := F(x,\ u(x),\ (u(x) - u(y))_{y \in X \setminus \{x\}}), \tag{5.10}$$

for all $u \in U$, $x \in X$. It is Discrete Degenerate Elliptic (DDE) iff F is non-decreasing w.r.t. the second argument $u(x)$, and w.r.t. each $u(x) - u(y)$, $y \in X \setminus \{x\}$.

Notation: the expression $Fu(x)$ should only be regarded as a shorthand for the accurate yet more verbose (5.10, right). In our application $X := \Omega_h$.

The numerical scheme $-\Delta_h^D$ is DDE on Ω_h, thanks to the non-negativity of the weights $(\rho_i)_{0 \le i \le 2}$, and to the finite difference expression (5.4, right) and (5.5), where u is extended to $\partial\Omega$ with the provided Dirichlet boundary values. On this basis we obtain a DDE discretization of nonlinear second order operators in Bellman form (5.1)

$$\Lambda_h u(x) := \sup_{\alpha \in \mathcal{A}} a_\alpha + b_\alpha u(x) - \Delta_h^{D_\alpha} u(x), \qquad \Lambda_h u(x) = \Lambda u(x) + \mathcal{O}(h^r), \tag{5.11}$$

where again $r = 1$ if x is adjacent to $\partial\Omega$, and $r = 2$ otherwise, at least if $\mathcal{A}$ is compact—which is the case for the Pucci operator. As shown in the next section, the supremum in (5.11, left) can

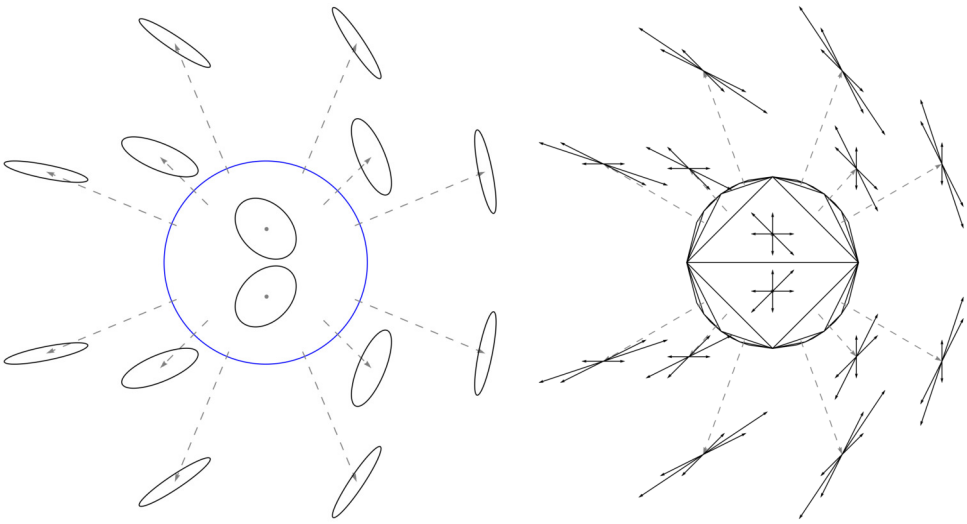


Figure 5.1: (Left) Ellipsoid $\{v \in \mathbb{R}^2;\, v^T \mathbf{D}(z) v = 1\}$ for some points z of the unit disc, see (5.9). Anisotropy degenerates as z moves toward the unit circle, shown blue. (Right) $\mathbf{D}(z)$-obtuse superbase, and opposites, for the same points z. This superbase is piecewise constant on an infinite triangulation of the unit disk [Sch09a], shown black.

be computed analytically in closed form, for the Pucci PDE, so that the numerical scheme Λ_h is explicit in terms of the unknown u.

Efficient construction of the Jacobian matrix of the numerical scheme. We use a Newton method to solve the discretized PDE, which requires assembling the sparse Jacobian matrix of the numerical scheme (5.11). In order to describe this essential step, let us rewrite the scheme in the following form (omitting the scale h for readability)

$$\max_{\alpha \in \mathcal{A}} \mathcal{F}(\alpha, x, u(x),\ (u(x) - u(y_1(x)))_{i=1}^{I}) = 0. \tag{5.12}$$

In comparison with (5.10), the expression (5.12) emphasizes (i) that F is defined as a maximum over a parameter set $\mathcal{A}$, and (ii) that the active stencil $y_1(x), \cdots, y_I(x)$ of a point $x \in \Omega_h$ only involves a small number of neighbors. The Jacobian matrix construction, at a given $u : \Omega_h \to \mathbb{R}$, involves the following steps:

1. Compute the maximizer $\alpha^*(x)$ in (5.12), for each $x \in \Omega_h$.
2. Differentiate the function $\mathcal{F}(\alpha^*(x), x, \delta,\ (\eta_i)_{i=1}^{I})$ w.r.t. parameters δ and $\eta_1, \cdots, \eta_I$, at the values $u(x)$ and $u(x) - u(y_i(x))$, $1 \le i \le I$, respectively.
3. Fill the corresponding entries of the sparse Jacobian matrix. More precisely, omitting the arguments of $\mathcal{F}$ for readability

$$J_{x,x} = \frac{\partial \mathcal{F}}{\partial \delta} + \sum_{1 \le i \le I} \frac{\partial \mathcal{F}}{\partial \eta_i}, \qquad J_{x,y_i(x)} = -\frac{\partial \mathcal{F}}{\partial \eta_i},\ 1 \le i \le I.$$

A custom automatic differentiation toolbox, open source and developed by Jean-Marie Mirebeau, makes these operations transparent. The above computations rely on the envelope theorem

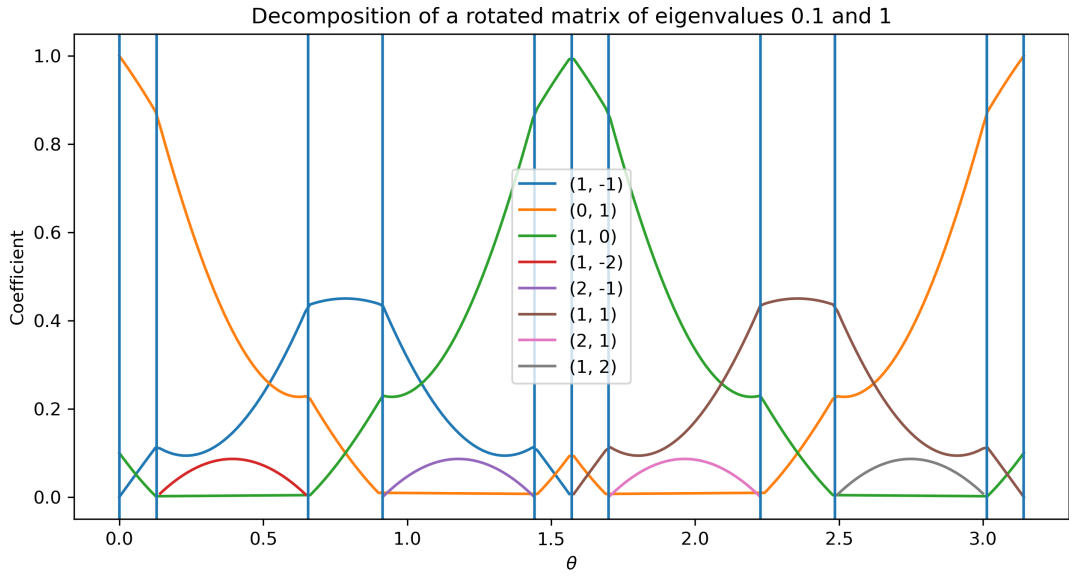


Figure 5.2: Coefficients of Selling's decomposition (5.7) of the matrix $D(\theta, \mu)$ for $\theta \in [0, \pi]$ and $\mu = 0.1$, see (5.13). The vertical bars correspond to the angles $0 = \theta_0 < \cdots < \theta_N = \pi$ where the support $e_0, e_1, e_2 \in \mathbb{Z}^2$ of the decomposition changes, see legend.

[Car01], which states that the value function to an optimization problem, here (5.12), over a compact set, here $\mathcal{A}$, is differentiable w.r.t. the parameters, here δ and $(\eta_i)_{i=1}^I$, whenever the problem solution, here $\alpha^*(x)$, is single valued (which is a generic property). In addition the first order derivatives have the expression used above, obtained by freezing the optimization parameter $\alpha \in \mathcal{A}$ to the optimal value $\alpha^*(x)$.

5.2.2 The Pucci operator

The Bellman form of the Pucci operator (5.3) involves a family of matrices $D(\theta, \mu)$, parameterized by the inverse $0 < \mu \leq 1$ of their condition number, and by an angle $0 \leq \theta \leq \pi$. As a starter, we rewrite those in the form (5.9)

$$D(\theta, \mu) = (1 + \mu)\, \mathbf{D}\Big(\frac{1 - \mu}{1 + \mu}\, \mathrm{n}(2\theta)\Big), \tag{5.13}$$

where $\mathrm{n}(\varphi) := (\cos\varphi, \sin\varphi)$. Note that the argument of $\mathbf{D}$ in (5.9) describes a circle of fixed radius $\frac{1-\mu}{1+\mu}$ within the unit disc, see Figure 5.1. Thus one can find $0 = \theta_0 < \cdots < \theta_N = \pi$, where $N = N(\mu)$, such that on each interval $[\theta_n, \theta_{n+1}]$ the superbase (e_0^n, e_1^n, e_2^n) is $D(\theta, \mu)$-obtuse and the coefficients in (5.7) take the form

$$\rho_i(\theta) = -\langle e_{i-1}^n, D(\theta, \mu) e_{i+1}^n \rangle = \alpha_i^n + \beta_i^n \cos(2\theta) + \gamma_i^n \sin(2\theta), \tag{5.14}$$

for suitable constants $\alpha_i^n, \beta_i^n, \gamma_i^n \in \mathbb{R}$, $0 \leq i \leq 2$, $0 \leq n < N$, see Figure 5.2. One finds that $N(1/4) = 2$, $N(1/10) = 10$, $N(1/400) = 122$, and one can show that $N(\mu) \leq C\mu^{-1}|\ln\mu|$ for some constant C independent of μ. By linearity of (5.8) one also has

$$\Delta_h^{D(\theta,\mu)} u(x) = \alpha^n + \beta^n \cos(2\theta) + \gamma^n \sin(2\theta) \tag{5.15}$$

for all $\theta \in [\theta_n, \theta_{n+1}]$, whose coefficients $\alpha^n, \beta^n, \gamma^n$ depend on ρ, h, u and x. Therefore, evaluating the discretized Bellman operator (5.11) associated with the Pucci equation (5.3) at a point $x \in \Omega_h$ amounts to solving a small number N of optimization problems, whose value is explicit. These optimization problems, and their value, take the following generic form

$$\max_{\varphi \in [\varphi_*, \varphi^*]} \alpha + \beta \cos\varphi + \gamma \sin\varphi$$
$$= \begin{cases} \alpha + \sqrt{\beta^2 + \gamma^2} & \text{if } \arg(\beta + i\gamma) \in]\varphi_*, \varphi^*[, \\ \alpha + \max\{\beta\cos\varphi_* + \gamma \sin\varphi_*,\ \beta\cos\varphi^* + \gamma\sin\varphi^*\} & \text{else,} \end{cases}$$

where $\arg(\omega)$ denotes the argument of $\omega \in \mathbb{C}$, taken in $[0, 2\pi[$. In view of (5.15), we choose $\varphi_* = 2\theta_n$, $\varphi^* = 2\theta_{n+1}$, $\alpha = \alpha^n$, $\beta = \beta^n$, and $\gamma = \gamma^n$. Then, following (5.3), we take the largest value among $0 \leq n < N$.

5.3 Numerical experiments

We present numerical results for the Pucci equation, chosen to illustrate the qualitative behavior of the solutions, and validate the scheme robustness and accuracy on synthetic problems with known solutions. Some of the considered domains are neither smooth nor convex, and the chosen synthetic solutions range from smooth to singular. The numerical scheme is implemented as described in the previous section, and a Newton method is used to solve the resulting coupled systems of non-linear equations. In practice, convergence to machine precision is achieved in a dozen of iterations, without damping, from an arbitrary guess. An open source Python® notebook reproducing (most of) the illustrations of this paper is available on Jean-Marie Mirebeau's webpage[1].

We illustrate on Figure 5.3 the transition of the Pucci equation from a strongly elliptic Laplacian-like PDE to a combinatorial-type convex-envelope problem, as the parameter μ takes values $1/4$ and $1/400$. The chosen domain is non-smooth and non-simply connected : $\Omega := U \setminus U'$ where $U := B(0,1) \cup (]0,1[\times]-1,1[)$ and $U' := kR_\theta(U)$ is its image under a scaling ($k = 0.4$) and a rotation ($\theta = \pi/3$). The boundary condition is 1 on ∂U, and 0 on $\partial U'$, and the r.h.s is $f \equiv 0$. The discretization grid size is 100×100, and the computation time is $1s$ for $\mu = 1/4$, and $45s$ for $\mu = 1/400$. The time difference is attributable to the complexity of the numerical scheme, which involves $N = 2$ pieces for in the first case and $N = 122$ in the latter, due to the larger condition number of the diffusion tensors $D(\theta, \mu)$, see §5.2.2. Nevertheless, the number $N = N(\mu)$ is independent of the grid scale, and both schemes are second order consistent. In the case $\mu = 1/400$, the PDE solution is quite close to the convex envelope of the boundary conditions, whose gradient is constant in some regions, and discontinuous across some lines, see Figure (5.3, right).

On figure 5.4, we reconstruct some known synthetic solutions from their image by the Pucci operator, with parameter $\mu = 0.2$, and their trace on the boundary. The examples are taken from the literature [FJ17; FO13], and the reconstruction errors are provided in the L^1 and L^∞ norm.

- (Smooth example [FJ17]) $u(x) = (x^2 + y^2)^2$ on $\Omega = B(0,1) \cup]0,1[^2$
- (C^1 example [FO13]) $u(x) = \max\{0, \|x - x_0\|^2 - 0.2)$ on $\Omega =]0,1[^2$.
- (Singular example [FO13]) $u(x) = \sqrt{2 - \|x\|^2}$ on $\Omega =]0,1[^2$.

Empirically, the L^1 numerical error behaves like $\mathcal{O}(h^2)$, where h is the grid scale (inverse of resolution in images). The L^∞ error behaves like $\mathcal{O}(h^2)$ in the smooth and C^1 examples, but

[1] Link : `Github.com/Mirebeau/AdaptiveGridDiscretizations`, see chapter 2.B.III.

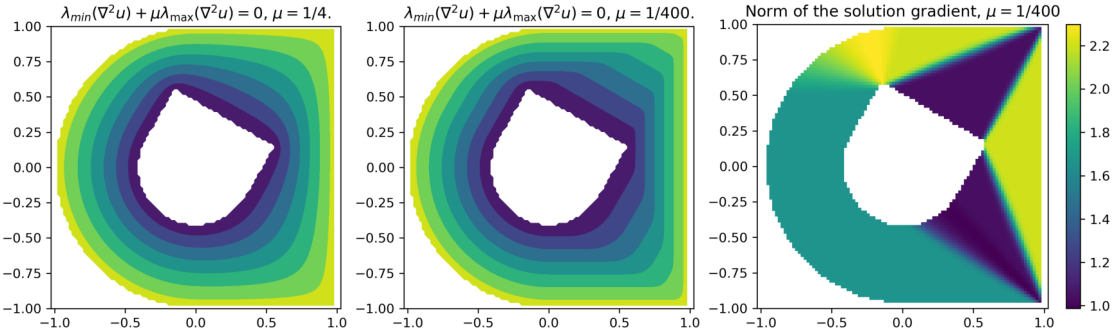


Figure 5.3: Solution of the Pucci PDE with $\mu = 1/4$ (left), $\mu = 1/400$ (center, right: gradient norm)

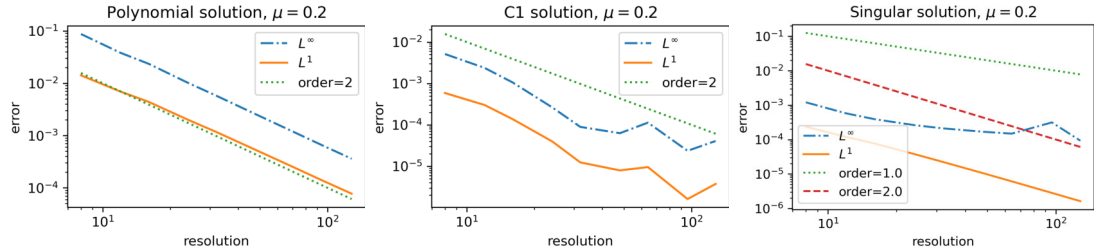


Figure 5.4: Numerical error as a function of grid size, for synthetic solutions to the Pucci equation.

decays more slowly for the singular solution. *Note: we rotated the Cartesian discretization grid by $\pi/3$ in these experiments, since otherwise the perfect alignment of the domain boundary with the coordinate axes gives an unfair advantage to grid based methods (like ours).*

5.4 Conclusion

In this paper, we presented a new strategy for discretizing non-divergence form, fully-nonlinear second order PDEs, and applied it to the Pucci equation. The steps of this approach can be summarized as follows: (i) rewrite the problem in Bellman form, as an extremum of linear equations, (ii) discretize the second order linear operators using monotone finite differences based on Selling's decomposition of positive definite matrices, (iii) solve the pointwise optimization problems involved in the numerical scheme definition, either explicitly (as could be done here), or numerically.

This methodology yields finite difference schemes which are degenerate elliptic, second order consistent, and use stencils of fixed size, in contrast with existing approaches [Obe08] which cannot achieve all these desirable properties simultaneously. Numerical experiments confirm that the proposed scheme can extract smooth PDE solutions with second order accuracy, and that it remains stable and convergent for harder problems involving a singularity at a point or along a line. Future research will be devoted to extending the results to other PDEs, such as the Monge-Ampère equation and its variants.

Chapter 6

Monotone discretization of the Monge-Ampère equation of optimal transport

This chapter corresponds to the paper [BM21].

6.1 Introduction

The problem of *optimal transport* [Vil09] is strongly related to the *Monge-Ampère equation* [GH09]: under suitable assumptions, the potential function which solves an optimal transport problem is also solution to the Monge-Ampère equation associated to this problem, equipped with the relevant boundary condition [DF14]. Some problems in nonimaging optics are also described by Monge-Ampère equations, among which some fit in the framework of optimal transport [CO08; GH09] and some do not [KO97; GH14].

Let us outline some approaches to the numerical resolution of optimal transport problems. One may solve an entropic regularization of a discrete optimal transport problem using Sinkhorn's iterations [Cut13]. The Benamou-Brenier method [BB00] is based on an extension of the optimal transport problem, with an added time variable. Some methods were also developed to solve semi-discrete optimal transport problems [KMT19], and applied to problems in nonimaging optics [DMT16]. Finally, one may solve numerically the Monge-Ampère equation associated to the considered optimal transport problem, as suggested in this paper and previously in [BD19; Fro19]. Benefits of this last approach include that it is easily adapted to various optimal transport problems by simply changing the coefficients of the approximated Monge-Ampère equation, and that one may use the theory of numerical schemes for degenerate elliptic partial differential equations [BS91] in order to establish convergence results.

We design a monotone finite difference discretization of the Monge-Ampère equation

$$\det{}_+ \left(D^2 u(x) - A(x, Du(x))\right) = B(x, Du(x)) \quad \text{in } X \tag{6.1}$$

where X is an open bounded subset of $\mathbb{R}^d$ containing the origin and A and B are bounded functions, whose values are respectively symmetric matrices and nonnegative numbers, A and $B^{1/d}$ being Lipschitz continuous with respect to their second variables uniformly with respect to their first variables, and A being continuous with respect to both its variables. For any symmetric

matrix M of size d, we denoted

$$\det\nolimits_+ M := \begin{cases} \det M & \text{if } M \succeq 0, \\ -\infty & \text{else.} \end{cases}$$

(We use the Loewner order on the space of symmetric matrices: $M_1 \succeq M_2$ if $M_1 - M_2$ is positive semidefinite. From now on, we denote respectively by $\mathcal{S}_d$, $\mathcal{S}_d^+$, and $\mathcal{S}_d^{++}$ the sets of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices of size d.)

Since we consider Monge-Ampère equations which are related to the problem of optimal transport, see section 6.5.1 and Remark 6.5.1, we also have to discretize the relevant boundary condition, described in section 6.1.2. We prove the *existence* of solutions, under suitable assumptions, to the proposed finite difference scheme. We also prove the *convergence* of solutions to this scheme, but only in the setting of *quadratic optimal transport*, where the function A is identically zero and the function B is separable in the form $B(x,p) = f(x)/g(p)$.

The Monge-Ampère equation is *degenerate elliptic*, meaning that it may be written in the form

$$F_{\mathrm{MA}}(x, Du(x), D^2u(x)) = 0 \quad \text{in } X, \tag{6.2}$$

where the operator $F_{\mathrm{MA}} \colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$ is *degenerate elliptic*, that is, nondecreasing with respect to its last variable: $F_{\mathrm{MA}}(x,p,M_1) \leq F_{\mathrm{MA}}(x,p,M_2)$ if $M_1 \succeq M_2$. The degenerate ellipticity property has a discrete counterpart which we call monotonicity, see Definition 6.2.5. Convergence of monotone schemes for degenerate elliptic equations may often be proved using a general argument, which was introduced in [BS91]. We use the fundamental part of this argument, see Theorem 6.2.7. As we discuss below Theorem 6.2.7, the full convergence result stated in [BS91] requires the approximated equation to satisfy a *strong comparison principle* which does not hold for the Monge-Ampère equation equipped with the boundary condition (6.22). Therefore, in order to prove Theorem 6.5.22, our convergence result in the setting of quadratic optimal transport, we need to establish an appropriate substitute to this comparison principle, in the form of Theorems 6.5.11 and 6.5.12.

One way to define the operator $F_{\mathrm{MA}}(x,p,M)$ so that it is both degenerate elliptic and consistent with (6.1) would be as

$$B(x,p) - \det\nolimits_+(M - A(x,p)). \tag{6.3}$$

This is not the definition we use, however. The reason is that there is no obvious way to build a monotone scheme by directly discretizing (6.3).

Instead, we use strategies described in [Lio85; Kry87] to reformulate the Monge-Ampère equation in the form (6.2), where F_{MA} is a supremum of semilinear operators (see also Proposition 6.5.8 for a more detailed description of what follows). First, note that formally, solutions to the Monge-Ampère equation satisfy the *admissibility* constraint

$$D^2u(x) \succeq A(x, Du(x)) \quad \text{in } X, \tag{6.4}$$

since otherwise the left-hand side in (6.1) would be equal to $-\infty$. For any symmetric positive semidefinite matrix M, it holds that

$$d(\det M)^{1/d} = \inf_{\substack{\mathcal{D} \in \mathcal{S}_d^{++} \\ \det \mathcal{D} = 1}} \langle \mathcal{D}, M \rangle = \inf_{\substack{\mathcal{D} \in \mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D}) = 1}} (\det \mathcal{D})^{-1/d} \langle \mathcal{D}, M \rangle, \tag{6.5}$$

where $\langle \mathcal{D}, M \rangle := \mathrm{Tr}(\mathcal{D}M)$. Choosing $M = D^2u(x) - A(x, Du(x))$ yields the two following reformulations of the Monge-Ampère equation (6.1):

$$B(x, Du(x)) - \inf_{\substack{\mathcal{D} \in \mathcal{S}_d^{++} \\ \det \mathcal{D} = 1}} \left(\frac{\langle \mathcal{D}, D^2u(x) - A(x, Du(x)) \rangle}{d} \right)^d = 0 \tag{6.6}$$

and alternatively, following [FJ17],

$$\max_{\substack{\mathcal{D} \in \mathcal{S}_d^{+} \\ \mathrm{Tr}(\mathcal{D}) = 1}} L_{\mathcal{D}} \left(B(x, Du(x)), D^2u(x) - A(x, Du(x)) \right) = 0 \quad \text{in } X, \tag{6.7}$$

where for any symmetric matrices $\mathcal{D}$ and M and nonnegative number b,

$$L_{\mathcal{D}}(b, M) := db^{1/d} (\det \mathcal{D})^{1/d} - \langle \mathcal{D}, M \rangle.$$

Note that the maximum in (6.7) is attained, as the maximum over a compact set of the continuous function $\mathcal{D} \mapsto L_{\mathcal{D}}(b, M)$ (this function is also concave, by the Minkowski determinant inequality). On the contrary, the parameter set of the infimum in (6.6) is not compact. Both reformulations enforce the admissibility constraint (6.4): for instance in (6.7), for any unit vector $e \in \mathbb{R}^d$, choosing $\mathcal{D} = e \otimes e$ in the maximum yields the inequality

$$\left\langle e, \left(D^2u(x) - A(x, Du(x)) \right) e \right\rangle \geq 0,$$

from which it follows that $D^2u(x) \succeq A(x, Du(x))$.

The numerical scheme that we study in this paper is a discretization of (6.7). Hence we define the operator F_{MA} in (6.2) by

$$F_{\mathrm{MA}}(x, p, M) := \max_{\substack{\mathcal{D} \in \mathcal{S}_d^{+} \\ \mathrm{Tr}(\mathcal{D}) = 1}} L_{\mathcal{D}}(B(x, p), M - A(x, p)). \tag{6.8}$$

6.1.1 Discretization of the Monge-Ampère equation

For any discretization step $h > 0$, we discretize the operator F_{MA} on a grid $\mathcal{G}_h \subset X \cap h\mathbb{Z}^d$. Denoting by d_H the Hausdorff distance between compact subsets of $\mathbb{R}^d$, which we recall is defined by

$$d_H(K_1, K_2) := \max \left\{ \max_{x \in K_1} \min_{y \in K_2} |x - y|, \max_{x \in K_2} \min_{y \in K_1} |x - y| \right\}, \tag{6.9}$$

we will assume that

$$\lim_{h \to 0} d_H \left(\partial X \cup ((X \cap h\mathbb{Z}^d) \setminus \mathcal{G}_h), \partial X \right) = 0, \tag{6.10}$$

or equivalently that if $K \subset X$ is compact, then for sufficiently small $h > 0$ one has $K \cap h\mathbb{Z}^2 \subset \mathcal{G}_h$. We will also need the technical assumption (6.39) of uniform connectedness of the grid $\mathcal{G}_h$.

Before introducing the discretization of F_{MA}, we need to define some finite difference operators. For any function $u \colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{Z}^d$, we define

$$T_h^e u[x] := \begin{cases} u[x + he] & \text{if } x + he \in \mathcal{G}_h, \\ +\infty & \text{else,} \end{cases} \tag{6.11}$$

$$\delta_h^e u[x] := \frac{T_h^e u[x] - u[x]}{h}, \qquad \Delta_h^e u[x] := \frac{T_h^e u[x] + T_h^{-e} u[x] - 2u[x]}{h^2}.$$

The constant $+\infty$ in the definition of T_h^e is related to the way we recommend discretizing the optimal transport boundary condition, discussed in section 6.1.2.

In the whole paper, we denote by $(e_1, \dots, e_d)$ the canonical basis of $\mathbb{Z}^d$. For any function $u\colon \mathcal{G}_h \to \mathbb{R}$ and point $x \in \mathcal{G}_h$, we define the Laplacian approximation and, whenever it makes sense, the centered gradient approximation

$$\Delta_h u[x] := \sum_{i=1}^d \Delta_h^{e_i} u[x], \qquad D_h u[x] := \left(\frac{\delta_h^{e_i} u[x] - \delta_h^{-e_i} u[x]}{2} \right)_{1 \le i \le d}. \tag{6.12}$$

We use Lax-Friedrichs approximations of the gradient of u in $A(x, Du(x))$ and $B(x, Du(x))$. To this end, we let $a_{\min} \le 0$, $a_{\mathrm{LF}} \ge 0$, and $b_{\mathrm{LF}} \ge 0$ be three constants independent of h. We will assume that for any $x \in \overline{X}$ and p, $p' \in \mathbb{R}^d$,

$$A(x,p) \succeq a_{\min} I_d, \tag{6.13}$$

$$|A(x,p) - A(x,p')|_2 \le a_{\mathrm{LF}} |p - p'|_1, \tag{6.14}$$

$$|B(x,p)^{1/d} - B(x,p')^{1/d}| \le b_{\mathrm{LF}} |p - p'|_1. \tag{6.15}$$

For any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{Z}^d$, we define

$$A_h^e u[x] := \begin{cases} a_{\min}|e|^2 \vee (\langle e, A(x, D_h u[x]) e \rangle - h a_{\mathrm{LF}} |e|^2 \Delta_h u[x]) & \text{if } \Delta_h u[x] < +\infty, \\ a_{\min}|e|^2 & \text{else,} \end{cases} \tag{6.16}$$

$$B_h u[x] := \begin{cases} 0 \vee (B(x, D_h u[x])^{1/d} - h b_{\mathrm{LF}} \Delta_h u[x])^d & \text{if } \Delta_h u[x] < +\infty, \\ 0 & \text{else.} \end{cases} \tag{6.17}$$

(In the whole paper, we denote respectively by $a \vee b$ and $a \wedge b$ the maximum and the minimum of two real numbers a and b.) For any family $v = (v_i)_{1 \le i \le I}$ of vectors of $\mathbb{Z}^d$ and any $\gamma \in \mathbb{R}^I$, we define

$$\mathcal{D}_v(\gamma) := \sum_{i=1}^I \gamma_i v_i \otimes v_i.$$

Finally, for any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and family v of vectors of $\mathbb{Z}^d$, we define

$$\Delta_h^v u[x] := (\Delta_h^e u[x])_{e \in v}, \qquad A_h^v u[x] := (A_h^e u[x])_{e \in v}.$$

For any $h > 0$, let V_h be a set of families of size $d(d+1)/2$ of vectors of $\mathbb{Z}^d$ such that

$$\lim_{h \to 0} d_H \left(\{ \mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}_+^{d(d+1)/2},\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1 \}, \{ \mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1 \} \right) = 0. \tag{6.18}$$

Equivalently, if $K \subset \mathcal{S}_d^{++}$ is compact, then for sufficiently small $h > 0$ each element of K can be written as $\mathcal{D}_v(\gamma)$ where $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$. We will also need to assume that

$$\lim_{h \to 0} h \max_{v \in V_h} \max_{e \in v} |e| = 0, \tag{6.19}$$

and that for any $h > 0$,

$$e_1 \in \bigcup_{v \in V_h} \bigcup_{e \in v} \{\pm e\}, \tag{6.20}$$

where we recall that e_1 denotes the first vector of the canonical basis of $\mathbb{R}^d$. We discretize F_{MA} by the operator $S^h_{\mathrm{MA}}\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ defined by

$$S^h_{\mathrm{MA}}u[x] := \max_{v\in V_h} \max_{\substack{\gamma\in\mathbb{R}_+^{d(d+1)/2}\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(B_h u[x], \Delta^v_h u[x] - A^v_h u[x]), \tag{6.21}$$

where for any family $v = (v_i)_{1\le i\le I}$ of vectors of $\mathbb{Z}^d$, $\gamma \in \mathbb{R}^I_+$, $b \ge 0$, and $m \in (\mathbb{R}\cup\{+\infty\})^I$,

$$L_{v,\gamma}(b,m) := db^{1/d}(\det \mathcal{D}_v(\gamma))^{1/d} - \langle \gamma, m\rangle.$$

Coefficients of γ are required to be nonnegative in order for the discretization to result in a numerical scheme which satisfies the monotonicity property (defined rigorously in Definition 6.2.12). Note that the constraint $\mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1$ may be rewritten as $\sum_{i=1}^{d(d+1)/2} \gamma_i |v_i|^2 = 1$.

In dimension $d = 2$, we recommend choosing V_h as a set of superbases of $\mathbb{Z}^2$:

Definition 6.1.1. A pair $v = (v_1, v_2)$ of vectors of $\mathbb{Z}^2$ is a *basis* of $\mathbb{Z}^2$ if $\det(v_1, v_2) = \pm 1$. A triple $v = (v_1, v_2, v_3)$ of vectors of $\mathbb{Z}^2$ is a *superbase* of $\mathbb{Z}^2$ if $v_1 + v_2 + v_3 = 0$ and $\det(v_1, v_2) = \pm 1$.

Note that in the definition above, the constraint $\det(v_1, v_2) = \pm 1$ is equivalent to $\det(v_2, v_3) = \pm 1$ or $\det(v_1, v_3) = \pm 1$. We explain in section 6.B how a set V_h of superbases of $\mathbb{Z}^2$ satisfying the above assumptions may be constructed. We prove in section 6.4 that when choosing V_h in this way, the second maximum in (6.21) admits a closed-form expression, at least when no infinite values are involved (infinite values may stem from the handling of the boundary condition, see (6.11), and a simple modification of the formula of Theorem 6.1.2 allows to compute the maximum in this case, by excluding finite differences whose value is infinite):

Theorem 6.1.2. *If $v = (v_1, v_2)$ is a basis of $\mathbb{Z}^2$, then for any $b \ge 0$ and $m \in \mathbb{R}^2$,*

$$\max_{\substack{\gamma\in\mathbb{R}_+^2\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \tilde{H}_v(b,m),$$

where

$$\tilde{H}_v(b,m) := \left(\frac{b}{|v_1|^2|v_2|^2} + \left(\frac{m_1}{2|v_1|^2} - \frac{m_2}{2|v_2|^2}\right)^2\right)^{1/2} - \frac{m_1}{2|v_1|^2} - \frac{m_2}{2|v_2|^2}.$$

If $v = (v_1, v_2, v_3)$ is a superbase of $\mathbb{Z}^2$, then for any $b \ge 0$ and $m \in \mathbb{R}^3$,

$$\max_{\substack{\gamma\in\mathbb{R}_+^3\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = H_v(b,m) \vee \max_{1\le i<j\le 3} \tilde{H}_{(v_i,v_j)}(b,m),$$

where

$$H_v(b,m) := \begin{cases} (b + \langle m, Q_v m\rangle)^{1/2} + \langle w_v, m\rangle & \text{if } Q_v m + (b + \langle m, Q_v m\rangle)^{1/2} w_v <_{\mathrm{vec}} 0,\\ -\infty & \text{else,}\end{cases}$$

$$Q_v := \frac{1}{4}\begin{pmatrix} |v_2|^2|v_3|^2 & \langle v_1, v_2\rangle |v_3|^2 & \langle v_1, v_3\rangle |v_2|^2\\ \langle v_1, v_2\rangle |v_3|^2 & |v_1|^2|v_3|^2 & \langle v_2, v_3\rangle |v_1|^2\\ \langle v_1, v_3\rangle |v_2|^2 & \langle v_2, v_3\rangle |v_1|^2 & |v_1|^2|v_2|^2\end{pmatrix}, \qquad w_v := \frac{1}{2}\begin{pmatrix}\langle v_2, v_3\rangle\\ \langle v_1, v_3\rangle\\ \langle v_1, v_2\rangle\end{pmatrix},$$

and, for $a \in \mathbb{R}^d$, we write $a <_{\mathrm{vec}} 0$ (respectively $a >_{\mathrm{vec}} 0$) if all components of a are negative (respectively positive).

6.1.2 Discretization of the boundary condition

In the setting of optimal transport, the relevant problem for the Monge-Ampère equation (6.1) is the *second boundary value problem*, which involves the *optimal transport boundary condition*

$$Du(x) \in \overline{P(x)}, \quad \forall x \in X, \tag{6.22}$$

where for any $x \in \overline{X}$, $P(x)$ is an open bounded convex nonempty subset of $\mathbb{R}^d$. We assume that $\overline{P(x)}$ depends continuously on x, for the Hausdorff distance d_H over compact subsets of $\mathbb{R}^d$ whose definition we recalled in (6.9). In the particular setting of quadratic optimal transport, in which we will prove convergence of the proposed numerical scheme, the set $P(x)$ does not depend on the variable x.

Note that despite being called a boundary condition, the constraint (6.22) involves the whole domain X. Some numerical approaches for solving the second boundary value problem, although not the one that we describe in this paper, rely on the fact that, in some cases, the constraint (6.22) can be reformulated in a way that only involves the boundary ∂X of the domain X, see for instance [BD19].

For now, let us consider the class of numerical schemes for equations (6.1) and (6.22) that are defined, for any discretization step $h > 0$, by an operator $S^h_{\mathrm{MABV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$, and may be written as

$$S^h_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{6.23}$$

One property of equations (6.1) and (6.22) is that their expressions depend only on derivatives of the function u and not on u itself, and therefore that the set of solutions is stable by addition of a constant. Accordingly, we say that the operator S^h_{MABV2} and the scheme (6.23) are *additively invariant* if for any function $u \colon \mathcal{G}_h \to \mathbb{R}$ and real number ξ, $S^h_{\mathrm{MABV2}}(u + \xi) = S^h_{\mathrm{MABV2}} u$.

We adapt the approach introduced in [Fro19] to build an operator S^h_{MABV2} suitable for (6.23). The idea is to build S^h_{MABV2} as a maximum of S^h_{MA} and of a monotone discretization $S^h_{\mathrm{BV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ of the left-hand side in a degenerate elliptic formulation of (6.22).

We use the following formulation of (6.22):

$$F_{\mathrm{BV2}}(x, Du(x)) \leq 0 \quad \text{in } X, \tag{6.24}$$

where $F_{\mathrm{BV2}} \colon \overline{X} \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$F_{\mathrm{BV2}}(x, p) := \max_{|e|=1} (\langle e, p \rangle - \sigma_{P(x)}(e)). \tag{6.25}$$

(We denote by $\sigma_{P(x)}$ the support function of the convex set $P(x)$: for any $e \in \mathbb{R}^d$, $\sigma_{P(x)}(e) := \sup_{p \in P(x)} \langle e, p \rangle$. Formally, if p belongs to the boundary $\partial P(x)$ of $P(x)$, then the maximum in the definition of F_{BV2} is attained when e is the unit outer normal of $\partial P(x)$ at point p.)

For any function $u \colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{R}^d$, we define the upwind finite difference

$$D^e_h u[x] := \sum_{i=1}^{d} ((0 \wedge \langle e, e_i \rangle) \delta^{e_i}_h u[x] - (0 \vee \langle e, e_i \rangle) \delta^{-e_i}_h u[x]),$$

using the convention $0 \times +\infty = 0$ (this convention is only needed in the immediate neighborhood of ∂X, where $\delta^{\pm e_i}_h u[x]$ may take infinite values). Then we define S^h_{BV2} and S^h_{MABV2} as

$$\begin{aligned} S^h_{\mathrm{BV2}} u[x] &:= \max_{|e|=1} (D^e_h u[x] - \sigma_{P(x)}(e)), \\ S^h_{\mathrm{MABV2}} u[x] &:= S^h_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \end{aligned} \tag{6.26}$$

In this setting, the scheme (6.23) is additively invariant.

Additively invariant schemes of the form (6.23) are not well-posed: their sets of solutions are stable by addition of a constant, thus not a singleton. Moreover they often have no solutions. One way to see this formally is that a well-posed scheme would need an additional equality to guarantee uniqueness of solutions, for instance $u[0] = 0$, but that then there would be one more equality than unknowns in the scheme. In the continuous setting, equations whose sets of solutions are stable by addition of a constant often admit solutions if and only if their coefficients satisfy some nonlocal condition, such as the mass balance condition (6.53) in the case of the Monge-Ampère equation of optimal transport; however, there may be no obvious discrete counterpart to this condition. See section 6.2 for further discussion of this issue.

In order to get around this difficulty, we solve an altered form of the scheme (6.23), following the approach used in the numerical experiments in [BD19]. We add an unknown α to the scheme, which must be a real number. For fixed α, we define the operators $S^{h,\alpha}_{\mathrm{MA}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ and $S^{h,\alpha}_{\mathrm{MABV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ as

$$S^{h,\alpha}_{\mathrm{MA}} u[x] := S^h_{\mathrm{MA}} u[x] + \alpha, \qquad S^{h,\alpha}_{\mathrm{MABV2}} u[x] := S^{h,\alpha}_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \tag{6.27}$$

The scheme we actually solve is

$$S^{h,\alpha}_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{6.28}$$

6.1.3 Main contributions and relation to previous works

We introduce the numerical scheme (6.28) for the Monge-Ampère equation (6.1), equipped with the boundary condition (6.22). We prove the existence of solutions to a class of monotone additively invariant numerical schemes featuring an additional unknown $\alpha \in \mathbb{R}$ as in (6.28), see section 6.2, and we show, in section 6.3, that the scheme (6.28) belongs to this class. This scheme is based on a discretization of the reformulation (6.7) of the Monge-Ampère equation. We prove in section 6.4 that this discretization admits a closed-form expression, as stated in Theorem 6.1.2. We prove convergence of the scheme in the setting of quadratic optimal transport, see section 6.5; convergence in the setting of more general optimal transport problems remains an open problem. We apply the scheme to the far field refractor problem in nonimaging optics, see section 6.6.

The closed-form expression obtained in Theorem 6.1.2 makes the implementation of the scheme particularly efficient, since no discretization of the parameter set of the maximum in (6.7) is needed. While to our knowledge the proposed discretization is the first one to admit such a closed-form expression among those that are based on the reformulation (6.7) of the Monge-Ampère equation, it is to be related to the MA-LBR scheme, introduced in [BCM16] in the setting of the Dirichlet problem for the Monge-Ampère equation when the function A is identically zero, and to the scheme we introduced in Chapter 5 for the Pucci equation. Both of the above-mentioned schemes involve the notion of superbases of $\mathbb{Z}^2$. We prove in section 6.A that the MA-LBR scheme is a discretization of (6.6), although it was not introduced as such in [BCM16].

As opposed to (6.6), the reformulation (6.7) has the benefit that its left-hand side remains finite even when (6.4) is not satisfied, and thus it is more stable numerically than (6.6). When solving schemes based on (6.6) using the damped Newton method, extremely small steps are typically required to ensure that the constraint (6.4) remains satisfied along the iterations; this is not the case with (6.7). Numerical schemes based on (6.7) were previously introduced in [FJ17], and then in [CWL18], although only in the setting of the Dirichlet problem for the Monge-Ampère equation when $A = 0$. In those papers, no counterpart of Theorem 6.1.2 was proved, hence the parameter set of the maximum in (6.7) had to be discretized.

Convergence of schemes for the second boundary value problem was previously studied in [BD19] and in [Fro19] in the setting of the quadratic optimal transport problem. Schemes considered in those two papers were based on the MA-LBR scheme introduced in [BCM16], and adapted in order to discretize the boundary condition (6.22).

In [BD19], convergence of a scheme of the form (6.23) was proved, but existence of solutions to this scheme was not. It turns out that solutions typically do not exist, due to the scheme being additively invariant. The approach used to solve the scheme in the numerical experiments was equivalent to adding an unknown $\alpha \in \mathbb{R}$ as in (6.28), but the proof of convergence was not extended to this setting.

In [Fro19], convergence of another scheme of the form (6.23) was proved. A Dirichlet boundary condition was enforced on ∂X, which in our setting would translate to replacing $+\infty$ by some constant $C \in \mathbb{R}$ in (6.11). Therefore the scheme considered in that paper is not additively invariant and does admit solutions. The Dirichlet boundary condition is to be understood in a weak sense (the one of viscosity solutions, see Definition 6.2.3), and may formally be simplified to $u(x) \leq C$ on ∂X, with equality at some point x_* of the boundary, provided that the scheme satisfied a property of *underestimation*, which is an assumption of the proof of convergence. This property is satisfied in the case of quadratic optimal transport at the cost of a careful handling of the constraint (6.22), but it does not seem obvious that it is satisfied for similar schemes in the case of more general optimal transport problems, with $A \neq 0$ in (6.1). No numerical experiments were performed in [Fro19]. In our experience, the scheme introduced in that paper has the drawback that the numerical error of its solutions tends to be unevenly distributed. This effect is related to the particular role played in the discretization by the point $x_* \in \partial X$ were the Dirichlet condition is satisfied in the classical sense, which leads to numerical artifacts and tends to decrease the accuracy of the scheme.

In our proof of convergence of the scheme (6.28), we use the arguments introduced in [Fro19] when appropriate. However, the property of underestimation is not required in our setting.

Note that the scheme (6.28), and its continuous counterpart (6.40) below, which both feature an additional unknown or parameter $\alpha \in \mathbb{R}$, fit in the framework of eigenvalue problems recently studied in [FL21]. Although our proof of convergence only applies to Monge-Ampère equation in the setting of quadratic optimal transport, our existence result, Theorem 6.2.14, is applicable to other such eigenvalue problems, as illustrated by the examples in section 6.2.

6.2 Monotone additively invariant schemes

6.2.1 Degenerate elliptic additively invariant equations

In this section, we study numerical schemes for a general degenerate elliptic equation of the form

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}. \tag{6.29}$$

Typically, F is discontinuous and $F(x, p, M)$ is defined differently depending on whether x belongs to X or to ∂X, in order to take into account the boundary condition in equation (6.29). The equation without the boundary condition would then be

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } X. \tag{6.30}$$

Let us recall the definition of degenerate ellipticity:

Definition 6.2.1 (Degenerate ellipticity)**.** The operator $F\colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$, and the equations (6.29) and (6.30), are *degenerate elliptic* if F is nonincreasing with respect to its last variable for the Loewner order: $F(x, p, M_1) \leq F(x, p, M_2)$ if $M_1 \succeq M_2$.

We say that equations (6.29) and (6.30) are *additively invariant* since, for reasonable notions of solutions, their sets of solutions are stable by addition of a constant, due to the fact that at any point x, the left-hand sides of those equations depend only on the derivatives $Du(x)$ and $D^2u(x)$ of the function u, and not on its value $u(x)$. This is not a standard property, and we will show it is a source of difficulty in the design of monotone numerical schemes. Typically, an additively invariant equation only has solutions if its coefficients are well-chosen and satisfy a particular nonlocal property.

Example 6.2.2. Throughout this section, we illustrate our definitions and results with Poisson's equation on the one-dimensional domain $X = (-1, 1)$, with the zero Neumann boundary condition:

$$\begin{cases} u''(x) = \psi(x) & \text{in } (-1,1), \\ u'(-1) = u'(1) = 0, \end{cases}$$

where $\psi\colon [-1,1] \to \mathbb{R}$ is an integrable function. We write this equation in the form

$$F_{\text{ex}}(x, u'(x), u''(x)) = 0 \quad \text{in } [-1,1], \tag{6.31}$$

where the degenerate elliptic operator $F_{\text{ex}}\colon [-1,1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$F_{\text{ex}}(x,p,m) := \begin{cases} -p & \text{if } x = -1, \\ p & \text{if } x = 1, \\ \psi(x) - m & \text{else.} \end{cases}$$

The equation only has solutions (respectively subsolutions, supersolutions) if $\int_{-1}^{1} \psi(x)\,dx = 0$ (respectively ≤ 0, ≥ 0), which we assume. Notice the similarity with the mass balance condition (6.53) which occurs in the setting of optimal transport.

An appropriate notion of solutions for degenerate elliptic equations, and for the study of discretizations of such equations, is the one of *viscosity solutions*. Before defining them, let us recall the definitions of the upper semicontinuous envelope F^* and lower semicontinuous envelope F_* of a function $F\colon E \to \overline{\mathbb{R}}$, E being a subset of $\mathbb{R}^n$: for any $x \in \overline{E}$,

$$F^*(x) := \limsup_{x' \to x} F(x), \qquad F_*(x) := \liminf_{x' \to x} F(x).$$

Definition 6.2.3 (Viscosity solution)**.** A function $u\colon \overline{X} \to \mathbb{R}$ is a *viscosity subsolution* to (6.29) if (i) it is upper semicontinuous and (ii) for any function φ in $C^2(\overline{X})$ and local maximum x of $u - \varphi$ in $\overline{X}$,

$$F_*(x, D\varphi(x), D^2\varphi(x)) \le 0.$$

It is a *viscosity supersolution* if (i) it is lower semicontinuous and (ii) for any function φ in $C^2(\overline{X})$ and local minimum x of $u - \varphi$ in $\overline{X}$,

$$F^*(x, D\varphi(x), D^2\varphi(x)) \ge 0.$$

It is a *viscosity solution* if it is both a viscosity subsolution and a viscosity supersolution. The same definitions, with $\overline{X}$ replaced by X, apply to equation (6.30).

Note that if a viscosity subsolution (respectively supersolution) u to (6.29) is twice differentiable at some point $x \in \overline{X}$ and if $F_*(x, Du(x), D^2u(x)) = F^*(x, Du(x), D^2u(x))$, then u is a classical subsolution (respectively supersolution) to (6.29) at point x.

6.2.2 Discretization

For any discretization step $h > 0$, let $\mathcal{G}_h$ be a finite nonempty subset of $\overline{X}$ containing the origin. In the rest of this paper, it is required that $\mathcal{G}_h$ be a subset of the Cartesian grid $X \cap h\mathbb{Z}^d$; however, this is not necessary in this section. What will be required in our definition of consistency is that

$$\lim_{h \to 0} d_H(\mathcal{G}_h, X) = 0. \tag{6.32}$$

Note that *in the case* that $\mathcal{G}_h$ is included in $X \cap h\mathbb{Z}^d$, then (6.32) is implied by (6.10).

We represent discretizations of the operator F by operators $S\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ that are additively invariant, according to the following definition:

Definition 6.2.4. An operator $S\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ is *additively invariant* if for any $u\colon \mathcal{G}_h \to \mathbb{R}$, $\xi \in \mathbb{R}$, and $x \in \mathcal{G}_h$, it holds that

$$S(u + \xi)[x] = Su[x].$$

For now, we let $S^h\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ be an additively invariant operator, for any $h > 0$, and we consider a numerical scheme of the form

$$S^h u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{6.33}$$

Definition 6.2.5. The scheme (6.33) is:

- *Monotone* if for any $h > 0$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $\overline{u}[x] = \underline{u}[x]$ and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h$, it holds that $S^h\overline{u}[x] \leq S^h\underline{u}[x]$.
- *Consistent* with equation (6.29) if (6.32) holds and for any $\varphi \in C^\infty(\overline{X})$ and $x \in \overline{X}$,

$$\limsup_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^h\varphi[x'] \leq F^*(x, D\varphi(x), D^2\varphi(x)),$$
$$\liminf_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^h\varphi[x'] \geq F_*(x, D\varphi(x), D^2\varphi(x)).$$

Remark 6.2.6. Schemes of the form (6.33) are typically called *degenerate elliptic* if for any $h > 0$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $\overline{u}[x] \leq \underline{u}[x]$ and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h \setminus \{x\}$, it holds that $S^h\overline{u}[x] \leq S^h\underline{u}[x]$. In our setting, monotonicity and degenerate ellipticity are equivalent, since operators S^h are additively invariant.

A framework is outlined in [BS91] for the proof of convergence of monotone schemes. The following fundamental result follows directly from the proof of [BS91, Theorem 2.1]:

Theorem 6.2.7. *Assume that there exist a sequence $(h_n)_{n\in\mathbb{N}}$ of discretization steps $h_n > 0$ converging to zero and a sequence $(u_n)_{n\in\mathbb{N}}$ of solutions $u_n\colon \mathcal{G}_{h_n} \to \mathbb{R}$ to (6.33) with $h = h_n$ such that $u_n[x]$ is bounded, uniformly over $n \in \mathbb{N}$ and $x \in \mathcal{G}_{h_n}$. If (6.33) is monotone and consistent with equation (6.29), then functions $\overline{u}$, $\underline{u}\colon \overline{X} \to \mathbb{R}$ defined by*

$$\overline{u}(x) := \limsup_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \qquad \underline{u}(x) := \liminf_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \tag{6.34}$$

are respectively a viscosity subsolution and supersolution to (6.29).

The definition of consistency in Definition 6.2.5 is slightly simpler than the one in [BS91], due to the assumption that operators S^h are additively invariant. In the framework of [BS91], in which the left-hand side in (6.29) may also depend on $u(x)$, a *strong comparison principle*, that is, a result stating that viscosity subsolutions to (6.29) are always less than viscosity supersolutions, is used after applying Theorem 6.2.7 to prove that $\overline{u} \le \underline{u}$, which allows to conclude that $\overline{u} = \underline{u}$, since $\overline{u} \ge \underline{u}$ by definition. Obviously, no strong comparison principle may hold if the set of viscosity solutions is nonempty and stable by addition of a constant. In our proof of convergence in the setting of quadratic optimal transport, we use Theorems 6.5.11 and 6.5.12 as a substitute to this comparison principle.

An important difficulty that we encounter is that numerical schemes of the form (6.33) typically have no solutions.

Example 6.2.8. Let $X = [-1, 1]$. For any $h > 0$, we let $\tilde{h} := \lceil h^{-1} \rceil^{-1}$, $\mathcal{G}_h := [-1, 1] \cap \tilde{h}\mathbb{Z}$, and we define the additively invariant operator $S^h_{\mathrm{ex}} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ by

$$S^h_{\mathrm{ex}} u[x] := \begin{cases} (u[-1] - u[-1 + \tilde{h}])/\tilde{h} & \text{if } x = -1, \\ (u[1] - u[1 - \tilde{h}])/\tilde{h} & \text{if } x = 1, \\ \psi(x) - (u[x + \tilde{h}] + u[x - \tilde{h}] - 2u[x])/\tilde{h}^2 & \text{else.} \end{cases}$$

Then the scheme

$$S^h_{\mathrm{ex}} u[x] = 0 \quad \text{in } \mathcal{G}_h$$

is monotone and consistent with equation (6.31). However, solving this scheme is equivalent to solving a square linear system, since the scheme operator $S^h_{\mathrm{ex}} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ is an affine map, and this linear system is noninvertible, since all constant functions belong to the kernel of the associated linear operator.

To get around this difficulty, we add a parameter $\alpha \in \mathbb{R}$ to the equation (6.29), yielding a new equation

$$F^\alpha(x, Du(x), D^2 u(x)) = 0 \quad \text{in } \overline{X}, \tag{6.35}$$

where for any $\alpha \in \mathbb{R}$, $F^\alpha \colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$ is a given operator, typically degenerate elliptic. The idea is to choose F^α so that $F^0 = F$ and (6.35) has no viscosity subsolutions when $\alpha > 0$ and no viscosity supersolutions when $\alpha < 0$.

Example 6.2.9. For any $\alpha \in \mathbb{R}$, we define $F^\alpha_{\mathrm{ex}} \colon [-1, 1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$F^\alpha_{\mathrm{ex}}(x, p, m) := \begin{cases} -p & \text{if } x = -1, \\ p & \text{if } x = 1, \\ \psi(x) - m + \alpha & \text{else.} \end{cases}$$

Then equation

$$F^\alpha_{\mathrm{ex}}(x, u'(x), u''(x)) = 0 \quad \text{in } \overline{X}$$

coincides with (6.31) when $\alpha = 0$, and only has solutions (respectively subsolutions, supersolutions) if $\int_{-1}^1 \psi(x)\, dx = -2\alpha$ (respectively $\le -2\alpha$, $\ge -2\alpha$). Recall that we assumed that $\int_{-1}^1 \psi(x)\, dx = 0$.

Accordingly, we add an unknown $\alpha \in \mathbb{R}$ to the numerical scheme. For any $h > 0$ and $\alpha \in \mathbb{R}$, we let $S^{h,\alpha} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ be an additively invariant operator, and we consider the scheme

$$S^{h,\alpha} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{6.36}$$

Example 6.2.10. In the setting of Example 6.2.8, for any $h > 0$ and $\alpha \in \mathbb{R}$, we define $S^{h,\alpha}_{\text{ex}} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ by

$$S^{h,\alpha}_{\text{ex}} u[x] := \begin{cases} (u[-1] - u[-1 + \tilde{h}])/\tilde{h} & \text{if } x = -1, \\ (u[1] - u[1 - \tilde{h}])/\tilde{h} & \text{if } x = 1, \\ \psi(x) - (u[x + \tilde{h}] + u[x - \tilde{h}] - 2u[x])/\tilde{h}^2 + \alpha & \text{else} \end{cases}$$

(recall that $\tilde{h} := \lceil h^{-1} \rceil^{-1}$). Then a solution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to the scheme

$$S^{h,\alpha}_{\text{ex}} u[x] = 0 \quad \text{in } \mathcal{G}_h$$

may easily be constructed explicitly.

The definition of solutions $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36) is obvious, but we will also need a notion of subsolutions (we could define supersolutions similarly, but this will not be needed):

Definition 6.2.11 (Subsolution)**.** Let $h > 0$. A pair $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ is a *subsolution* to (6.36) if $S^{h,\alpha} u[x] \leq 0$ in $\mathcal{G}_h$.

Since α is an unknown of the scheme, and not simply a fixed parameter, Definition 6.2.5 needs to be adapted to this new setting. We also define some other properties that the scheme (6.36) may satisfy. Conceptually, the following definition is intended for schemes such that $S^{h,\alpha} u[x]$ is non-decreasing with respect to α.

Definition 6.2.12. The scheme (6.36) is:

- *Monotone* if for any $\alpha \in \mathbb{R}$, the scheme (6.33) with $S^h = S^{h,\alpha}$ is monotone in the sense of Definition 6.2.5.
- *Consistent* with the parametrized equation (6.35) if for any family of real numbers $(\alpha_h)_{h>0}$ converging to some $\alpha \in \mathbb{R}$ as h approaches zero, the scheme (6.33) with $S^h = S^{h,\alpha_h}$ is consistent with equation (6.35) in the sense of Definition 6.2.5.
- *Continuous* if for any small $h > 0$, the map $\mathbb{R} \times \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$, $(\alpha, u) \mapsto S^{h,\alpha} u$ takes finite values and is continuous.
- *Stable* if the following properties hold:
  - (i) For any small $h > 0$, there exists a subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36).
  - (ii) There exists a nonincreasing function $\omega \colon \mathbb{R} \to \mathbb{R}_+$ such that for any small $h > 0$, any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36), and any x_1, $x_2 \in \mathcal{G}_h$, one has
  $$|u[x_1] - u[x_2]| \leq \omega(\alpha).$$
  - (iii) There exists $\alpha_0 \in \mathbb{R}$ such that for any small $h > 0$ and any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36), one has $\alpha \leq \alpha_0$.
  - (iv) There exists $\alpha_1 \in \mathbb{R}$ such that for any small $h > 0$ and any solution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36), one has $\alpha \geq \alpha_1$.
- *Equicontinuously stable* if it satisfies all items in the definition of stability above, with (ii) replaced by the following:
  - (ii') There exists a function $\omega \colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, nonincreasing with respect to its first variable and satisfying $\lim_{t \to 0} \omega(\alpha, t) = 0$ for any $\alpha \in \mathbb{R}$, such that for any small $h > 0$, any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (6.36), and any x_1, $x_2 \in \mathcal{G}_h$, one has
  $$|u[x_1] - u[x_2]| \leq \omega(\alpha, |x_1 - x_2|).$$

Note that the value $\alpha = 0$ does not play a special role in Definition 6.2.12. The role of the functions ω in the definitions of stability and equicontinuous stability is to allow schemes to become unstable when $\alpha \to -\infty$.

Obviously, if (6.36) is equicontinuously stable, then it is stable. In the case of the scheme considered in this paper for the Monge-Ampère equation, subsolutions will be established to be uniformly Lipschitz continuous, which is stronger than equicontinuity, see the proof of Proposition 6.3.6. In particular, the boundary condition $u(x) - \infty = 0$ on ∂X (to be understood in the viscosity sense, as mentioned in section 6.1) does not induce a boundary layer.

Theorem 6.2.7 is easily adapted to the scheme (6.36):

Corollary 6.2.13. *Assume that there exist a sequence $(h_n)_{n\in\mathbb{N}}$ of discretization steps $h_n > 0$ converging to zero, a sequence $(\alpha_n)_{n\in\mathbb{N}}$ of real numbers α_n converging to some $\alpha \in \mathbb{R}$, and a sequence $(u_n)_{n\in\mathbb{N}}$ of functions $u_n \colon \mathcal{G}_{h_n} \to \mathbb{R}$ such that (α_n, u_n) is solution to (6.36) with $h = h_n$ and $u_n[x]$ is bounded, uniformly over $n \in \mathbb{N}$ and $x \in \mathcal{G}_{h_n}$. If (6.36) is monotone and consistent with (6.35), then limits superior and inferior $\overline{u}$, $\underline{u} \colon \overline{X} \to \mathbb{R}$ defined as in (6.34) are respectively a viscosity subsolution and supersolution to (6.35) in $\overline{X}$.*

If (6.36) is equicontinuously stable, then Corollary 6.2.13 is simplified by the fact that, by the Arzelà-Ascoli theorem, sequences $(\alpha_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ converge uniformly, up to extracting a subsequence, to some $\alpha \in \mathbb{R}$, and to some continuous function $u \colon \overline{X} \to \mathbb{R}$, which coincides with the limits superior and inferior $\overline{u}$ and $\underline{u}$ for this subsequence.

6.2.3 Existence

Our main result in this section concerns existence of solutions to the scheme (6.36). The proof is an adaptation of Perron's method to this setting. While we assume that (6.36) is stable in the sense of Definition 6.2.12, this assumption may be relaxed, see Remark 6.2.15 below.

Theorem 6.2.14 (Existence)**.** *Assume that (6.36) is monotone, continuous, and stable. Then for small $h > 0$, there exists a solution to (6.36).*

Proof. We define the set

$$U := \{(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h} \mid S^{h,\alpha}u[x] \leq 0 \text{ in } \mathcal{G}_h\}$$

of subsolutions to (6.36). Since we assumed that (6.36) is stable, U is nonempty and there exists $\alpha \in \mathbb{R}$ defined by

$$\alpha := \sup_{(\overline{\alpha},\overline{u})\in U} \overline{\alpha}. \tag{6.37}$$

Let us show that there exists $u \colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (6.36). Let $((\alpha_n, u_n))_{n\in\mathbb{N}}$ be a maximizing sequence in the definition of α, and let $\alpha_* := \min_{n\in\mathbb{N}} \alpha_n$. We may assume, up to adding a constant to u_n, that $u_n[0] = 0$ for any $n \in \mathbb{N}$. Then by stability, $|u_n[x]| = |u_n[x] - u_n[0]| \leq \omega(\alpha_n) \leq \omega(\alpha_*)$, for any $n \in \mathbb{N}$ and $x \in \mathbb{R}^{\mathcal{G}_h}$. This means that the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}^{\mathcal{G}_h}$ and thus that it converges, up to extracting a subsequence, to some function $\hat{u} \colon \mathcal{G}_h \to \mathbb{R}$. By continuity of the scheme, $(\alpha, \hat{u})$, as the limit of subsolutions $((\alpha_n, u_n))_{n\in\mathbb{N}}$, is a subsolution to (6.36).

Among all functions $u \colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (6.36), we choose one which maximizes the cardinal of the set $\mathcal{G}_* := \{x \in \mathcal{G}_h \mid S^{h,\alpha}u[x] < 0\}$. Let us show how such a function u may be transformed into a solution to the scheme.

First note that $\mathcal{G}_*$ may not be equal to $\mathcal{G}_h$, since in this case, by continuity of the scheme, there would exist $\alpha' > \alpha$ close enough to α so that $(\alpha', u) \in U$, contradicting (6.37).

Knowing that $\mathcal{G}_* \neq \mathcal{G}_h$, and using stability, we may define, for small $\varepsilon > 0$, the function $\tilde{u}_\varepsilon \colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}_\varepsilon[x] := \sup\{\overline{u}[x] \mid (\alpha, \overline{u}) \in U,\ \overline{u} = u \text{ in } \mathcal{G}_h \setminus \mathcal{G}_*,\ S^{h,\alpha}\overline{u}[x] \leq -\varepsilon \text{ in } \mathcal{G}_*\}. \tag{6.38}$$

To ensure that the supremum above is the one of a nonempty set, we choose ε small enough so that u itself is suitable choice of function $\overline{u}$. By continuity of the scheme, we may pass to the limit in maximizing sequences and deduce that for any $x \in \mathcal{G}_h$, there exists $\overline{u} \colon \mathcal{G}_h \to \mathbb{R}$ such that $(\alpha, \overline{u}) \in U$, $S^{h,\alpha}\overline{u}[x] \leq -\varepsilon$ in $\mathcal{G}_*$, $\tilde{u}_\varepsilon \geq \overline{u}$ in $\mathcal{G}_h$, and $\tilde{u}_\varepsilon[x] = \overline{u}[x]$. Then by monotonicity, $S^{h,\alpha}\tilde{u}_\varepsilon[x] \leq S^{h,\alpha}\overline{u}[x]$. It follows that $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (6.36) and that $S^{h,\alpha}\tilde{u}_\varepsilon[x] \leq -\varepsilon$ in $\mathcal{G}_*$.

Let us show that $S^{h,\alpha}\tilde{u}_\varepsilon[x] = -\varepsilon$ in $\mathcal{G}_*$. Assume that there exists $x_* \in \mathcal{G}_*$ so that $S^{h,\alpha}\tilde{u}_\varepsilon[x_*] < -\varepsilon$. For any $\delta > 0$, we define $\tilde{u}_{\varepsilon,\delta} \colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}_{\varepsilon,\delta}[x] := \begin{cases} \tilde{u}_\varepsilon[x] + \delta & \text{if } x = x_*, \\ \tilde{u}_\varepsilon[x] & \text{else.} \end{cases}$$

By monotonicity, $S^{h,\alpha}\tilde{u}_{\varepsilon,\delta}[x] \leq S^{h,\alpha}\tilde{u}_\varepsilon[x]$ for any $x \in \mathcal{G}_h \setminus \{x_*\}$, and by continuity, we may choose δ small enough so that $S^{h,\alpha}\tilde{u}_{\varepsilon,\delta}[x_*] \leq -\varepsilon$. This contradicts (6.38), since $\tilde{u}_{\varepsilon,\delta}$ is a suitable choice for $\overline{u}$ and $\tilde{u}_{\varepsilon,\delta}[x_*] > \tilde{u}_\varepsilon[x_*]$.

We now define $\tilde{u} \colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}[x] := \lim_{\varepsilon \to 0} \tilde{u}_\varepsilon[x].$$

Note that the right-hand side is the limit of a bounded nondecreasing sequence. By continuity, $S^{h,\alpha}\tilde{u}[x] = 0$ in $\mathcal{G}_*$ and $(\alpha, \tilde{u})$ is a subsolution to (6.36). Let us show that it is a solution. If it is not the case, then there exists $x_* \in \mathcal{G}_h \setminus \mathcal{G}_*$ such that $S^{h,\alpha}\tilde{u}[x_*] < 0$. By continuity, there exists $\varepsilon > 0$ such that $S^{h,\alpha}\tilde{u}_\varepsilon[x_*] < 0$. Since $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (6.36) and $S^{h,\alpha}\tilde{u}_\varepsilon[x] < 0$ in $\mathcal{G}_*$, this contradicts the assumption that $\mathcal{G}_*$ is of maximal cardinal. Thus $(\alpha, \tilde{u})$ is necessarily a solution to (6.36). □

Remark 6.2.15. Since $h > 0$ is fixed in Theorem 6.2.14, the subsolution, the function ω, and the number α_0 in (i), (ii), and (iii) in the definition of stability of the scheme (Definition 6.2.12) only need to exist for this fixed value of h. Also, (iv) is not needed.

6.3 Properties of the proposed scheme

In this section, we show that the scheme (6.28) satisfies the properties we defined in section 6.2. First note that for any $h > 0$ and $\alpha \in \mathbb{R}$, the operator $S^{h,\alpha}_{\mathrm{MABV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ is additively invariant.

Proposition 6.3.1 (Monotonicity)**.** *Assume the Lipschitz regularity properties* (6.14) *and* (6.15)*. Then the scheme* (6.28) *is monotone, in the sense of Definition 6.2.12.*

Proof. Let $h > 0$, $\alpha \in \mathbb{R}$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u} \colon \mathcal{G}_h \to \mathbb{R}$ be such that $\overline{u}[x] = \underline{u}[x]$ and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h$. We need to show that

$$S^{h,\alpha}_{\mathrm{MABV2}}\overline{u}[x] \leq S^{h,\alpha}_{\mathrm{MABV2}}\underline{u}[x].$$

By the definition (6.27) of the operator $S^{h,\alpha}_{\mathrm{MABV2}}$, it suffices to prove that both $S^h_{\mathrm{MA}}\overline{u}[x] \leq S^h_{\mathrm{MA}}\underline{u}[x]$ and $S^h_{\mathrm{BV2}}\overline{u}[x] \leq S^h_{\mathrm{BV2}}\underline{u}[x]$. The second inequality follows directly from the definition (6.26) of S^h_{BV2}, so let us prove the first one.

By the definition (6.21) of S^h_{MA}, it suffices to prove that for any family $v = (v_1, \dots, v_I)$ of vectors of $\mathbb{Z}^d$ and any $\gamma \in \mathbb{R}^I_+$,

$$L_{v,\gamma}(B_h\overline{u}[x], \Delta^v_h\overline{u}[x] - A^v_h\overline{u}[x]) \le L_{v,\gamma}(B_h\underline{u}[x], \Delta^v_h\underline{u}[x] - A^v_h\underline{u}[x]).$$

First note that the operator Δ^v_h was defined so that $\Delta^v_h\overline{u}[x] \ge \Delta^v_h\underline{u}[x]$ elementwise. If $B_h\overline{u}[x] = 0$, then $B_h\overline{u}[x]^{1/d} \le B_h\underline{u}[x]^{1/d}$, since B^h is a nonnegative operator. If $B_h\overline{u}[x] > 0$ (which, by definition of B_h, implies that $x \pm he_i \in \mathcal{G}_h$ for any $i \in \{1, \dots, d\}$), then, using (6.15) for the second inequality,

$$\begin{aligned}
B_h\overline{u}[x]^{1/d} - B_h\underline{u}[x]^{1/d} &\le B(x, D_h\overline{u}[x])^{1/d} - B(x, D_h\underline{u}[x])^{1/d} - hb_{\mathrm{LF}}\Delta_h(\overline{u} - \underline{u})[x] \\
&\le b_{\mathrm{LF}}(|D_h\overline{u}[x] - D_h\underline{u}[x]|_1 - h\Delta_h(\overline{u} - \underline{u})[x]) \\
&= \frac{b_{\mathrm{LF}}}{h}\sum_{i=1}^{d}\Big(|(\overline{u} - \underline{u})[x + he_i] - (\overline{u} - \underline{u})[x - he_i]| \\
&\qquad\qquad - (\overline{u} - \underline{u})[x + he_i] - (\overline{u} - \underline{u})[x - he_i]\Big) \\
&\le 0,
\end{aligned}$$

and thus $B_h\overline{u}[x]^{1/d} \le B_h\underline{u}[x]^{1/d}$. Similarly, for any $e \in v$, if $A^e_h\overline{u}[x] = a_{\min}|e|^2$, then $A^e_h\overline{u}[x] \le A^e_h\underline{u}[x]$, and otherwise, using (6.14),

$$\begin{aligned}
A^e_h\overline{u}[x] - A^e_h\underline{u}[x] &\le \langle e, (A(x, D_h\overline{u}[x]) - A(x, D_h\underline{u}[x]))e\rangle - ha_{\mathrm{LF}}|e|^2\Delta_h(\overline{u} - \underline{u})[x] \\
&\le a_{\mathrm{LF}}|e|^2(|D_h\overline{u}[x] - D_h\underline{u}[x]|_1 - \Delta_h(\overline{u} - \underline{u})[x]) \le 0,
\end{aligned}$$

hence $A^e_h\overline{u}[x] \le A^e_h\underline{u}[x]$. We easily conclude that $S^h_{\mathrm{MA}}\overline{u}[x] \le S^h_{\mathrm{MA}}\underline{u}[x]$. □

From the grid $\mathcal{G}_h$, we may build a graph whose nodes are the points of $\mathcal{G}_h$ and whose edges are pairs of points that are neighbors on the grid, that is, between whom the Euclidean distance is equal to h. To prove other properties of the scheme, we need the technical assumption that the distance on this graph, multiplied by h, is equivalent to the Euclidean distance, uniformly over small $h > 0$. Equivalently, we require that there exists some positive constant $C_{\mathcal{G}}$, such that for any small $h > 0$ and any function $\varphi\colon \mathcal{G}_h \to \mathbb{R}$,

$$\max_{\substack{x_1, x_2 \in \mathcal{G}_h \\ x_1 \ne x_2}} \frac{|\varphi[x_1] - \varphi[x_2]|}{|x_1 - x_2|} \le C_{\mathcal{G}} \max_{\substack{x_1, x_2 \in \mathcal{G}_h \\ |x_1 - x_2| = h}} \frac{|\varphi[x_1] - \varphi[x_2]|}{h}. \tag{6.39}$$

Proposition 6.3.2 (Continuity)**.** *Assume* (6.39)*. Then the scheme* (6.28) *is continuous, in the sense of Definition 6.2.12.*

Proof. For any $x \in \mathcal{G}_h$, the function $\mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}$, $u \mapsto S^{h,\alpha}_{\mathrm{MABV2}}u[x]$ is a maximum over a compact set of continuous functions with values in $\mathbb{R} \cup \{-\infty\}$, see (6.21), (6.26), and (6.27). Hence it is a continuous function with values in $\mathbb{R} \cup \{-\infty\}$. It remains to prove that $S^{h,\alpha}_{\mathrm{MABV2}}u[x] > -\infty$.

By (6.39), there exists $e = \pm e_i$, $i \in \{1, \dots, d\}$, such that $x - he \in \mathcal{G}_h$. Therefore

$$S^{h,\alpha}_{\mathrm{MABV2}}u[x] \ge S^h_{\mathrm{BV2}}u[x] \ge D^e_h u[x] - \sigma_{P(x)}(e) = -\delta^{-e}_h u[x] - \sigma_{P(x)}(e) > -\infty,$$

which concludes the proof. □

Let us now study consistency of the scheme (6.28) with the degenerate elliptic equation

$$F^{\alpha}_{\mathrm{MABV2}}(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}, \tag{6.40}$$

where for any $\alpha \in \mathbb{R}$, $x \in \overline{X}$, $p \in \mathbb{R}^d$, and $M \in \mathcal{S}_d$,

$$F^{\alpha}_{\mathrm{MABV2}}(x,p,M) := \begin{cases} (F_{\mathrm{MA}}(x,p,M) + \alpha) \vee F_{\mathrm{BV2}}(x,p) & \text{if } x \in X, \\ -\infty & \text{else,} \end{cases}$$

and $F_{\mathrm{MA}}(x,p,M)$ and $F_{\mathrm{BV2}}(x,p)$ are defined respectively in (6.8) and (6.25). We first prove a consistency property that is stronger to the one we introduced in Definition 6.2.12, and that will be useful in the study of stability of the scheme.

Proposition 6.3.3 (Consistency)**.** *Assume* (6.10)*,* (6.13)*,* (6.18)*, and* (6.19)*. Let $\varphi \in C^\infty(\overline{X})$ and $(\alpha_h)_{h>0}$ be a family of real numbers converging to some $\alpha \in \mathbb{R}$ as h approaches zero. Then*

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \le F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{6.41}$$

uniformly over $x \in \mathcal{G}_h$ and $\alpha \in \mathbb{R}$. Moreover, for any compact subset K of X,

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \ge F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{6.42}$$

uniformly over $x \in K \cap \mathcal{G}_h$ and $\alpha \in \mathbb{R}$.

Proof. Let K be a compact subset of X. For convenience, when $a_h(x)$ and $b_h(x)$ are real numbers depending on $h > 0$ and on $x \in \mathcal{G}_h$, we write $a_h(x) \leqq_K b_h(x)$ if $a_h(x) \le b_h(x)$ for any $h > 0$ and $x \in \mathcal{G}_h$, with equality if $x \in K$. Then it suffices to show that

$$S^h_{\mathrm{MA}}\varphi[x] \leqq_K F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{6.43}$$

$$S^h_{\mathrm{BV2}}\varphi[x] \leqq_K F_{\mathrm{BV2}}(x, D\varphi(x)) + o_{h\to 0}(1), \tag{6.44}$$

uniformly over $x \in \mathcal{G}_h$.

For any $x \in \mathcal{G}_h$ and $i \in \{1, \dots, d\}$, it holds that $T_h^{\pm e_i}\varphi[x] \ge \varphi(x \pm he_i)$, and using (6.10), we may assume that h is small enough so that the equality $T_h^{\pm e_i}\varphi[x] = \varphi(x \pm he_i)$ holds whenever $x \in K$. Then injecting first-order Taylor expansions of φ in the definition of S^h_{BV2} yields (6.44).

If $x \in \mathcal{G}_h$ is such that $\Delta_h\varphi[x] < +\infty$, then $x \pm he_i \in \mathcal{G}_h$ for any $i \in \{1, \dots, d\}$, and thus $D_h\varphi[x] = D\varphi(x) + O(h^2)$ and $\Delta_h\varphi[x] = \Delta\varphi(x) + O(h^2)$. In particular, $\Delta_h\varphi[x]$ is bounded. Therefore, using that B is Lipschitz continuous with respect to its last variable, uniformly with respect to its first variable,

$$B(x, D_h\varphi[x])^{1/d} - hb_{\mathrm{LF}}\Delta_h\varphi[x] = B(x, D\varphi(x))^{1/d} + O(h).$$

Since $B \ge 0$ and using the definition (6.17) of B_h, it follows that

$$B_h\varphi[x]^{1/d} = B(x, D\varphi(x))^{1/d} + O(h).$$

Now if $\Delta_h\varphi[x] = +\infty$ (by (6.10), for h small, this may only happen if $x \notin K$), it holds that $B_h\varphi[x] = 0 \le B(x, D\varphi(x))$. We deduce that

$$B_h\varphi[x]^{1/d} \leqq_K B(x, D\varphi(x))^{1/d} + O(h)$$

uniformly over $x \in \mathcal{G}_h$. Similarly, for any $v \in V_h$ and $e \in v$, we may assume, using (6.10) and (6.19), that h is small enough so that $x \pm he \in \mathcal{G}_h$ whenever $x \in K \cap \mathcal{G}_h$, and then, using (6.13) and the same reasoning as above,

$$A^e_h\varphi[x] \leqq_K \langle e, A(x, D\varphi(x))e\rangle + O(h|e|^2),$$

$$-\Delta_h^e\varphi[x] \leqq_K -\langle e, D^2\varphi(x)e\rangle + O(h^2|e|^4),$$

uniformly over $x \in \mathcal{G}_h$. Then for any $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$ such that $\mathrm{Tr}(\mathcal{D}_v(\gamma)) = \sum_{i=1}^{d(d+1)/2} \gamma_i |v_i|^2 = 1$, using (6.19) for the last equality,

$$\begin{aligned}
-\langle \gamma, \Delta_h^v\varphi[x] - A_h^v\varphi[x]\rangle &= -\sum_{i=1}^{d(d+1)/2} \gamma_i(\Delta_h^{v_i}\varphi[x] - A_h^{v_i}\varphi[x]) \\
&\leqq_K -\sum_{i=1}^{d(d+1)/2} \gamma_i \left\langle v_i, \left(D^2\varphi(x) - A(x, D\varphi(x))\right) v_i\right\rangle \\
&\quad + \sum_{i=1}^{d(d+1)/2} \gamma_i O(h|v_i|^2 + h^2|v_i|^4) \\
&= -\langle \mathcal{D}_v(\gamma), D^2\varphi(x) - A(x, D\varphi(x))\rangle + O(h + h^2|v_i|^2) \\
&= -\langle \mathcal{D}_v(\gamma), D^2\varphi[x] - A(x, D\varphi(x))\rangle + o_{h\to 0}(1),
\end{aligned} \tag{6.45}$$

uniformly over $x \in \mathcal{G}_h$, v, and γ. Thus

$$S_{\mathrm{MA}}^h\varphi[x] \leqq_K \max_{v\in V_h} \max_{\substack{\gamma\in\mathbb{R}_+^{d(d+1)/2} \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{\mathcal{D}_v(\gamma)}\left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) + o_{h\to 0}(1).$$

We deduce (6.43) using (6.18) and that the affine map

$$\{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\} \to \mathbb{R}, \quad \mathcal{D} \mapsto L_{\mathcal{D}}(b, M) \tag{6.46}$$

is continuous, uniformly over b and M belonging to compact sets. □

Remark 6.3.4 (Order of consistency). Under appropriate assumptions, the order of consistency of the scheme (6.28) is easily deduced from the proof of Proposition 6.3.3. Let $\varphi \in C^\infty(X)$, and let $K \subset X$ be compact. Then, for small $h > 0$ and uniformly over $x \in K \cap \mathcal{G}_h$,

$$S_{\mathrm{BV2}}^h\varphi[x] = F_{\mathrm{BV2}}(x, D\varphi(x)) + O(h).$$

For the operator S_{MA}^h, we distinguish two cases:

(*General case*) If there exist $r_1 > 0$ and $r_2 \in (0, 1)$ such that the following refinements of (6.18) and (6.19) hold:

$$\begin{aligned}
&d_H\left(\{\mathcal{D}_v(\gamma) \mid v \in V_h, \gamma \in \mathbb{R}_+^{d(d+1)/2}, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}, \{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\}\right) = O(h^{r_1}), \\
&\max_{v\in V_h} \max_{e\in v} |e| = O(h^{-r_2}),
\end{aligned}$$

then, refining the last equality in (6.45) and using that the map (6.46) is $1/d$-Hölder continuous, one has, for small $h > 0$ and uniformly over $x \in K \cap \mathcal{G}_h$,

$$S_{\mathrm{MA}}^h\varphi[x] = F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) + O(h^{1\wedge(2-2r_2)\wedge(r_1/d)}).$$

In dimension $d = 2$, when choosing V_h as in Remark 6.B.9, one has $r_1 = 2r$ and $r_2 = r$, hence S_{MA}^h is consistent with F_{MA} to the order $1 \wedge (2 - 2r) \wedge r$, and the optimal choice for r is $r = 2/3$, yielding consistency to the order $2/3$.

(*Smooth case*) The consistency is improved if (6.2) admits a solution $u \in C^2(\overline{X})$ such that, uniformly over K, $D^2u(x) - A(x, Du(x)) \in \mathcal{S}_d^{++}$ has condition number less than some constant $c > 1$. In this setting, the maximum in (6.8) is attained for $\mathcal{D} = (D^2u(x) - A(x, Du(x)))^{-1} / \operatorname{Tr}((D^2u(x) - A(x, Du(x)))^{-1})$, which has condition number less than c for $x \in K$. We recommend choosing the set V_h independently of h, but such that any $\mathcal{D} \in \mathcal{S}_d^{++}$ with condition number less than c is of the form $\mathcal{D} = \mathcal{D}_v(\gamma)$ for some $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$ (see section 6.B for a suitable construction of V_h in dimension $d = 2$). Then (6.18) is not satisfied, but in a neighborhood of the solution u, the operator S^h_{MA} is still consistent with F_{MA}, to the order one, uniformly over $x \in K$.

In practice, one may choose to implement the scheme with Lax-Friedrichs relaxation parameters $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$, as we do in section 6.6. The drawback of doing this is that (6.14) and (6.15), and thus Proposition 6.3.1, do not hold anymore unless $A(x, p)$ and $B(x, p)$ do not depend on p. The benefit is that consistency is improved. In the setting of the smooth case described above, if $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$, then, in a neighborhood of u and uniformly over $x \in K$, S^h_{MA} is consistent with F_{MA} to the order two.

Note that the order of consistency of the whole scheme (6.28) is the minimum of the ones of S^h_{BV2} and S^h_{MA}, but for a fixed point x, the order is the one of the operator for which the maximum is reached in (6.27), which in practice is $S^{h,\alpha}_{\mathrm{MA}} = S^h_{\mathrm{MA}} + \alpha$ at most points of the grid.

Corollary 6.3.5 (Consistency)**.** *Assume* (6.10)*,* (6.13)*,* (6.18)*, and* (6.19)*. Then the scheme* (6.28) *is consistent with equation* (6.40)*, in the sense of Definition 6.2.12.*

Proof. We have to show that if φ, $(\alpha_h)_{h>0}$, and α are as in Proposition 6.3.3, then for any $x \in \overline{X}$,

$$\limsup_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \le (F^{\alpha}_{\mathrm{MABV2}})^*(x, D\varphi(x), D^2\varphi(x)), \tag{6.47}$$

$$\liminf_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \ge (F^{\alpha}_{\mathrm{MABV2}})_*(x, D\varphi(x), D^2\varphi(x)). \tag{6.48}$$

If $x \in X$, then (6.47) and (6.48) follow respectively from (6.41) and (6.42), taking first the limit over h and then the limit over x'. If $x \in \partial X$, then (6.47) follows from (6.41) and (6.48) is always true, since $(F^{\alpha}_{\mathrm{MABV2}})_*(x, D\varphi(x), D^2\varphi(x)) = -\infty$. □

Finally, we establish stability of the proposed scheme.

Proposition 6.3.6 (Equicontinuous stability)**.** *Assume* (6.10)*,* (6.13) *to* (6.15)*,* (6.18) *to* (6.20)*, and* (6.39)*. If there exists a function $\varphi \in C^\infty(\overline{X})$ such that for any $x \in \overline{X}$, $D\varphi(x) \in P(x)$, then the scheme* (6.28) *is equicontinuously stable, in the sense of Definition 6.2.12.*

Proof. Let us check all items in the definition of equicontinuous stability.

(i) The function φ was chosen so that $F_{\mathrm{BV2}}(x, D\varphi(x)) < 0$, uniformly over $x \in \overline{X}$. Also, since A and B are bounded, there exists $\alpha_1 \le 0$ such that $F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) < -\alpha_1$, uniformly over $x \in \overline{X}$. It follows that $(F^{\alpha_1}_{\mathrm{MABV2}})^*(x, D\varphi(x), D^2\varphi(x)) < 0$, uniformly over $x \in \overline{X}$. Then by Proposition 6.3.3, for any small $h > 0$ and any $x \in \mathcal{G}_h$, $S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x] < 0$. Hence (α_1, φ) is a subsolution to (6.28) for small $h > 0$.

(ii') Let $h > 0$ be small and let $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ be a subsolution to (6.28). Then for any $x \in \mathcal{G}_h$, $S^h_{\mathrm{BV2}}u[x] \le 0$. Choosing $e = \pm e_i$, $i \in \{1, \dots, d\}$ in the definition of S^h_{BV2}, it follows that $-\delta_h^{\pm e_i}u[x] \le \sigma_{P(x)}(\mp e_i)$. Since the compact set $\overline{P(x)}$ is continuous with respect to $x \in \overline{X}$ for the Hausdorff distance, there exists $C_P \ge 0$ such that for any $x \in \overline{X}$ and $i \in \{1, \dots, d\}$, $\sigma_{P(x)}(\pm e_i) \le C_P$. Hence $-\delta_h^{\pm e_i}u[x] \le C_P$. Using (6.39), we easily deduce that

$$\max_{\substack{x_1, x_2 \in \mathcal{G}_h \\ x_1 \neq x_2}} \frac{|u[x_1] - u[x_2]|}{|x_1 - x_2|} \le C_{\mathcal{G}} C_P.$$

Hence (ii') holds with $\omega(\alpha,t) := C_{\mathcal{G}} C_P t$.

(iii) Let $h>0$ be small and $(\alpha,u)\in\mathbb{R}\times\mathbb{R}^{\mathcal{G}_h}$ be a subsolution to (6.28). Then for any $x\in\mathcal{G}_h$, $S^h_{\mathrm{MA}}u[x]\le -\alpha$. By (6.20), there exists $v\in V_h$ and $\gamma\in\mathbb{R}_+^{d(d+1)/2}$ such that $\mathcal{D}_v(\gamma)=e_1\otimes e_1$ (and thus $\mathrm{Tr}(\mathcal{D}_v(\gamma))=1$). Choosing v and γ as parameters in the definition of S^h_{MA} yields $A^{e_1}_h u[x]-\Delta^{e_1}_h u[x]\le -\alpha$. Since $A^{e_1}_h u[x]\ge a_{\min}$, it follows that $\Delta^{e_1}_h u[x]\ge a_{\min}+\alpha$.

Let $\ell>0$, independent of h, be such that the segment $[0,\ell e_1]$ belongs to X (recall that $0\in X$ by assumption), and let $n_h := \lceil \ell/h\rceil$. By (6.10), we may assume that h is small enough so that $ihe_1\in X$, for any $i\in\{0,\dots,n_h+1\}$. Then for any $i\in\{1,\dots,n_h\}$, $h\Delta^{e_1}_h u[ihe_1]=\delta^{e_1}_h u[ihe_1]+\delta^{-e_1}_h u[ihe_1]=\delta^{e_1}_h u[ihe_1]-\delta^{e_1}_h u[(i-1)he_1]$, hence $\delta^{e_1}_h u[ihe_1]=\delta^{e_1}_h u[(i-1)he_1]+h\Delta^{e_1}_h u[ihe_1]$ and

$$\delta^{e_1}_h u[n_h h e_1]=\delta^{e_1}_h u[0]+h\sum_{i=1}^{n_h}\Delta^{e_1}_h u[ihe_1]\ge \delta^{e_1}_h u[0]+n_h h(a_{\min}+\alpha).$$

Since $n_h h\ge \ell$, if $\alpha\ge -a_{\min}$, then

$$\delta^{e_1}_h u[n_h h e_1]\ge \delta^{e_1}_h u[0]+\ell(a_{\min}+\alpha)=-\delta^{-e_1}_h u[he_1]+\ell(a_{\min}+\alpha).$$

We proved in (ii) that $\delta^{e_1}_h u[n_h h e_1]\le C_P$ and $\delta^{-e_1}_h u[he_1]\le C_P$. Therefore

$$\alpha\le \frac{2C_P}{\ell}-a_{\min}.$$

(iv) Let $h>0$ be small and $(\alpha,u)\in\mathbb{R}\times\mathbb{R}^{\mathcal{G}_h}$ be a solution to (6.28) (note that in the proof, we only use that it is a supersolution). Let $\alpha_1\ge 0$ be as in (i). Up to adding a constant to u, we may assume that there exists $x\in\mathcal{G}_h$ such that $u[x]=\varphi[x]$ and $u\ge\varphi$ in $\mathcal{G}_h$. Then by Proposition 6.3.1, $S^{h,\alpha_1}_{\mathrm{MABV2}}u[x]\le S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x]$. We proved in (i) that $S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x]<0$. Thus $S^{h,\alpha_1}_{\mathrm{MABV2}}u[x]<0$, and by definition of $S^{h,\alpha_1}_{\mathrm{MABV2}}$, it holds that $S^h_{\mathrm{BV2}}u[x]<0$ and $S^h_{\mathrm{MA}}u[x]<-\alpha_1$. On the other hand, the equality $S^{h,\alpha}_{\mathrm{MABV2}}u[x]=0$ may be expanded as

$$S^h_{\mathrm{BV2}}u[x]\vee(S^h_{\mathrm{MA}}u[x]+\alpha)=0.$$

Since $S^h_{\mathrm{BV2}}u[x]<0$, we deduce that $\alpha=-S^h_{\mathrm{MA}}u[x]>\alpha_1$. □

Note that in the proof of item (ii'), we actually proved that solutions to the scheme are Lipschitz continuous uniformly over small $h>0$.

The existence of a suitable function φ in Proposition 6.3.6 is a natural assumption in the setting of optimal transport. We defer discussion of this assumption to section 6.5.1, and in particular to Remark 6.5.1.

6.4 Closed-form formula in dimension two

This section is devoted to the proof of Theorem 6.1.2.

Remark 6.4.1 (Numerical complexity of the scheme). The motivation for Theorem 6.1.2 is to improve the numerical efficiency of the scheme.

Consider a two-dimensional Cartesian grid $\mathcal{G}_h$ with $O(N^2)$ points. Assume that at any point $x\in\mathcal{G}_h$, one has to perform respectively M_{MA} and M_{BV2} operations in order to compute $S^h_{\mathrm{MA}}u[x]$ and $S^h_{\mathrm{BV2}}u[x]$. Then the overall numerical complexity of the scheme on the grid $\mathcal{G}_h$ is $O(N^2(M_{\mathrm{MA}}+M_{\mathrm{BV2}}))$.

When using Theorem 6.1.2 in the implementation of the scheme, M_{MA} is proportional to the number of superbases in the set V_h. As in Remark 6.3.4, we distinguish between the *smooth case*

and the *general case*. In the smooth case, V_h does not depend on N, hence $M_{\mathrm{MA}} = O(1)$. In the general case, if V_h is built as in Remark 6.B.9, with $r = 2/3$ as suggested by Remark 6.3.4, then by Proposition 6.B.10, $M_{\mathrm{MA}} = O(N^{2/3} \log N)$.

For comparison, one could choose to discretize the parameter set of the maximum in the definition (6.8) of the operator S^h_{MA} instead of using Theorem 6.1.2, and in this case M_{MA} would be proportional to the number of points in this discretization. Since the set of symmetric positive semidefinite matrices of size two and of unit trace has dimension two, in order to guarantee consistency of the scheme to some order $r > 0$, one should choose at least $M_{\mathrm{MA}} = O(N^{2r})$. This is more costly than using Theorem 6.1.2, both in the smooth case (in which the desired order, according to Remark 6.3.4, is $r = 1$, or even $r = 2$ if $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$) and in the general case (in which the desired order is $r = 2/3$).

There is also a maximum in the definition (6.26) of S^h_{BV2} which, depending on the expression of the set-valued function P in (6.22), either admits a closed-form formula or needs to be discretized. If it admits a closed-form formula, then M_{BV2} does not depend on N. If it needs to be discretized, then M_{BV2} is proportional to the number of points in the discretization and, in order to guarantee consistency of the operator S^h_{BV2} with F_{BV2} at some order $r > 0$, one should choose $M_{\mathrm{BV2}} = O(N^r)$, since the parameter set is one-dimensional. The numerical cost of this discretization is negligible in the general case, but not in the smooth case. In practice, in many applications, the maximum in (6.27) is only attained by $S^h_{\mathrm{BV2}}u[x]$ at points $x \in \mathcal{G}_h$ that are close to ∂X. A perspective for future research would be to prove that one may use a variant of the scheme (6.28) which would only require computing $S^h_{\mathrm{BV2}}u[x]$ at such points, reducing the numerical cost of handling the boundary condition (6.22).

In dimension $d = 2$, choosing V^h as a family of superbases of $\mathbb{Z}^2$ (see Definition 6.1.1) is motivated by *Selling's formula* [Sel74]: for any family $v = (v_1, v_2, v_3)$ of vectors of $\mathbb{Z}^2$, recall that we defined $\gamma \colon \mathbb{R}^3 \to \mathcal{S}_2^+$ by

$$\mathcal{D}_v(\gamma) := \sum_{i=1}^{3} \gamma_i v_i \otimes v_i,$$

and let us also define $\gamma_v \colon \mathcal{S}_2 \to \mathbb{R}^3$ by

$$\gamma_v(\mathcal{D}) := (-\langle v_{i+1}^\perp, \mathcal{D} v_{i+2}^\perp \rangle)_{1 \le i \le 3}, \tag{6.49}$$

where we consider the indices of the elements of v modulo three, and where if $e = (a, b) \in \mathbb{R}^2$, we denote $e^\perp := (-b, a)$.

Proposition 6.4.2 (Selling's formula)**.** *If $v = (v_1, v_2, v_3)$ is a superbase of $\mathbb{Z}^2$, then γ_v is the inverse bijection of $\mathcal{D}_v$: for any $\mathcal{D} \in \mathcal{S}_2$, $\mathcal{D} = \mathcal{D}_v(\gamma_v(\mathcal{D}))$.*

Proof. It suffices to show that for any $1 \le i \le j \le 2$,

$$\langle v_i^\perp, \mathcal{D} v_j^\perp \rangle = \langle v_i^\perp, \mathcal{D}_v(\gamma_v(\mathcal{D})) v_j^\perp \rangle.$$

This is easily verified using the properties of superbases of $\mathbb{Z}^2$ and the fact that for any $\{i, j\} \subset \{1, 2, 3\}$, $\langle v_i^\perp, v_j \rangle = \det(v_i, v_j)$. □

Proof of Theorem 6.1.2. We prove separately the two statements of the theorem.

Case of bases. Let $v = (v_1, v_2)$ be a basis of $\mathbb{Z}^2$, $b \ge 0$, and $m = (m_1, m_2) \in \mathbb{R}^2$. Note that

$$\{\gamma \in \mathbb{R}_2^+ \mid \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\} = \left\{ \left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2} \right) \mid t \in [-1, 1] \right\},$$

as the segment whose endpoints are $(1/|v_1|^2, 0)$ and $(0, 1/|v_2|^2)$. Then

$$\begin{aligned}&\max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m)\\ &= \max_{t\in[-1,1]} \left(2b^{1/2}\left(\det \mathcal{D}_v\left(\left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2}\right)\right)\right)^{1/2} - \frac{1+t}{2|v_1|^2}m_1 - \frac{1-t}{2|v_2|^2}m_2\right).\end{aligned}$$

We compute that for any $t \in [-1,1]$,

$$\begin{aligned}\det \mathcal{D}_v\left(\left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2}\right)\right) &= \det\left(\frac{(1+t)}{2|v_1|^2} v_1\otimes v_1 + \frac{(1-t)}{2|v_2|^2} v_2\otimes v_2\right)\\ &= \frac{1}{4}(1-t^2)\frac{\det(v_1,v_2)^2}{|v_1|^2|v_2|^2} = \frac{(1-t^2)}{4|v_1|^2|v_2|^2},\end{aligned}$$

using the definition of $\mathcal{D}_v$ for the first equality, that $\det(a\otimes a + b\otimes b) = \det(a,b)^2$ for any a, $b \in \mathbb{R}^2$ for the second equality, and that $\det(v_1, v_2) = \pm 1$ for the third equality. After defining $\omega_v^{(0)} \in \mathbb{R}$ and $\omega_v^{(1)}$, $\omega_v^{(2)} \in \mathbb{R}^2$ by

$$\omega_v^{(0)} := \frac{1}{|v_1|^2|v_2|^2}, \qquad \omega_v^{(1)} := \frac{1}{2}\begin{pmatrix} 1/|v_1|^2 \\ -1/|v_2|^2\end{pmatrix}, \qquad \omega_v^{(2)} := \frac{1}{2}\begin{pmatrix} 1/|v_1|^2 \\ 1/|v_2|^2\end{pmatrix},$$

it follows that

$$\max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{t\in[-1,1]} \left((\omega_v^{(0)})^{1/2} b^{1/2}(1-t^2)^{1/2} - \langle\omega_v^{(1)}, m\rangle t - \langle\omega_v^{(2)}, m\rangle\right).$$

This is the maximum of a concave function over $[-1,1]$. Writing the first order optimality condition yields that the optimal t must satisfy

$$t^2 = \frac{\langle\omega_v^{(1)}, m\rangle^2}{\omega_v^{(0)} b + \langle\omega_v^{(1)}, m\rangle^2},$$

from which we deduce the expected formula

$$\max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = (\omega_v^{(0)} b + \langle\omega_v^{(1)}, m\rangle^2)^{1/2} - \langle\omega_v^{(2)}, m\rangle = \tilde{H}_v(b,m).$$

Case of superbases. We use that in the space of symmetric matrices size two equipped with the Frobenius norm, the set of symmetric positive semidefinite matrices of unit trace is a disk. More precisely, let us define the affine map $\mathfrak{D}\colon \mathbb{R}^2 \to \mathcal{S}_2$ by

$$\mathfrak{D}(\rho) = \frac{1}{2}\begin{pmatrix} 1+\rho_1 & \rho_2 \\ \rho_2 & 1-\rho_1\end{pmatrix}. \tag{6.50}$$

Note that the above definition is closely related to Pauli matrices in quantum mechanics. It is easily proved that

$$\{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\} = \{\mathfrak{D}(\rho) \mid |\rho| \le 1\}. \tag{6.51}$$

Moreover, for any $\rho \in \mathbb{R}^d$ such that $|\rho| \le 1$,

$$\det \mathfrak{D}(\rho) = \frac{1}{4}(1-|\rho|^2), \qquad \mathrm{Cond}(\mathfrak{D}(\rho)) = \frac{1+|\rho|}{1-|\rho|}. \tag{6.52}$$

Let $v = (v_1, v_2, v_3)$ be a superbase of $\mathbb{Z}^2$, $b \geq 0$, and $m \in \mathbb{R}^3$. The Minkowski determinant inequality states, in any dimension $d \in \mathbb{N}$, the function $\det(\cdot)^{1/d}$ is concave over $\mathcal{S}_d^+$. Hence the function

$$\{\gamma \in \mathbb{R}^3 \mid \mathcal{D}_v(\gamma) \succeq 0,\ \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\} \to \mathbb{R}, \quad \gamma \mapsto L_{v,\gamma}(b,m)$$

is concave too. Recall that $\mathcal{D}_v(\gamma) \succeq 0$ whenever $\gamma \in \mathbb{R}^3_+$. Let

$$\gamma_v^*(b,m) \in \operatorname*{argmax}_{\substack{\gamma \in \mathbb{R}^3 \\ \mathcal{D}_v(\gamma) \succeq 0 \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m).$$

If the strict elementwise inequality $\gamma_v^*(b,m) >_{\mathrm{vec}} 0$ is not satisfied, then

$$\max_{\substack{\gamma \in \mathbb{R}^3_+ \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{1 \leq i < j \leq 3} \max_{\substack{\gamma \in \mathbb{R}^2_+ \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{(v_i,v_j),\gamma}(b,m) = \max_{1 \leq i < j \leq 3} \tilde{H}_{(v_i,v_j)}(b,m),$$

since the maximum in the left-hand side is attained on the boundary of the parameter set. Thus it suffices to prove that

$$H_v(b,m) = \begin{cases} L_{v,\gamma_v^*(b,m)}(b,m) & \text{if } \gamma_v^*(b,m) >_{\mathrm{vec}} 0, \\ -\infty & \text{else.} \end{cases}$$

Let us prove the above. If $\gamma_v \colon \mathcal{S}_2 \to \mathbb{R}^3$ and $\mathfrak{D} \colon \mathbb{R}^2 \to \mathcal{S}_2$ are functions defined respectively by (6.49) and (6.50), then, by (6.51) and Selling's Formula (Proposition 6.4.2), it holds that

$$\max_{\substack{\gamma \in \mathbb{R}^3 \\ \mathcal{D}_v(\gamma) \succeq 0 \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{|\rho| \leq 1} L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m),$$

and there exists

$$\rho_v^*(b,m) \in \operatorname*{argmax}_{|\rho| \leq 1} L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m)$$

such that

$$\gamma_v^*(b,m) = \gamma_v(\mathfrak{D}(\rho_v^*(b,m))).$$

Let

$$W_v := \frac{1}{2} \begin{pmatrix} v_{2,1}v_{3,1} - v_{2,2}v_{3,2} & v_{2,1}v_{3,2} + v_{2,2}v_{3,1} \\ v_{1,1}v_{3,1} - v_{1,2}v_{3,2} & v_{1,1}v_{3,2} + v_{1,2}v_{3,1} \\ v_{1,1}v_{2,1} - v_{1,2}v_{2,2} & v_{1,1}v_{2,2} + v_{1,2}v_{2,1} \end{pmatrix}.$$

Recall that $Q_v \in \mathcal{S}_3$ and $w_v \in \mathbb{R}^3$ were defined in the statement of the theorem, and note that $Q_v = W_v W_v^\top$. It is easily computed that for any $\rho \in \mathbb{R}^2$,

$$\gamma_v(\mathfrak{D}(\rho)) = W_v \rho - w_v,$$

and thus, using also (6.52), that

$$L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m) = b^{1/2}(1 - |\rho|^2)^{1/2} - \langle W_v \rho - w_v, m \rangle.$$

Therefore, $\rho_v^*(b,m)$ is the argmax of a concave function over the unit disk, and writing the first-order optimality condition yields

$$\rho_v^*(v,m) = -\frac{W_v^\top m}{(b + |W_v^\top m|^2)^{1/2}} = -\frac{W_v^\top m}{(b + \langle m, Q_v m \rangle)^{1/2}}.$$

Thus

$$\gamma_v^*(b,m) = \gamma_v(\mathfrak{D}(\rho_v^*(b,m))) = -\frac{Q_v m}{(b+\langle m, Q_v m\rangle)^{1/2}} - w_v$$

and

$$L_{v,\gamma_v^*(b,m)}(b,m) = L_{v,\gamma_v(\mathfrak{D}(\rho_v^*(b,m)))}(b,m) = (b+\langle m, Q_v m\rangle)^{1/2} + \langle w_v, m\rangle,$$

which concludes the proof. □

6.5 Application to quadratic optimal transport

6.5.1 The quadratic optimal transport problem

Let Y be an open bounded convex nonempty subset of $\mathbb{R}^d$ and $f\colon \overline{X} \to \mathbb{R}_+$ and $g\colon \overline{Y} \to \mathbb{R}_+^*$ be two densities satisfying the mass balance condition

$$\int_X f(x)\,dx = \int_Y g(y)\,dy, \tag{6.53}$$

f being bounded and continuous almost everywhere and $1/g$ being Lipschitz continuous. For convenience, in this paper we extend the function g to the whole domain $\mathbb{R}^d$ in such a manner that $1/g\colon \mathbb{R}^d \to \mathbb{R}_+^*$ is bounded and Lipschitz continuous.

In the *quadratic optimal transport problem* between f and g, one aims to solve the minimization problem

$$\inf_{T_\# f = g} \int_X |x - T(x)|^2 f(x)\,dx, \tag{6.54}$$

where the unknown is a Borel map $T\colon X \to \overline{Y}$ and the constraint $T_\# f = g$ means that for any Borel subset E of Y,

$$\int_{T^{-1}(E)} f(x)\,dx = \int_E g(y)\,dy. \tag{6.55}$$

In the literature, it is typically assumed that:

$$\text{the set } X \text{ is convex.} \tag{6.56}$$

For simplicity, we will sometimes assume instead that:

$$\text{the set } X \text{ is strongly convex.} \tag{6.57}$$

It was proved in [Bre91] (see also [Vil03, Theorem 2.12]) that, under assumption (6.56), the optimal transport problem (6.54) admits a solution T which is the gradient of a convex function $u\colon X \to \mathbb{R}$, called the *potential function* of the problem. Then, if u is smooth enough, it may be deduced by performing the change of variables $y = T(x)$ in the right-hand side of (6.55) that u is solution to the Monge-Ampère equation (6.1), where

$$A(x,p) = 0, \qquad B(x,p) = \frac{f(x)}{g(p)}. \tag{6.58}$$

Additionally, the constraint that $T(x) = Du(x) \in \overline{Y}$, for any $x \in X$, may be written as (6.22), where for any $x \in \overline{X}$,

$$P(x) = Y. \tag{6.59}$$

Note that in this setting, a possible choice of function φ in Proposition 6.3.6 is given by $\varphi(x) := \langle x, y_0\rangle$, for some $y_0 \in Y$.

Remark 6.5.1 (General optimal transport). In the general optimal transport problem, a cost function $c \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$ is given, and one aims to solve

$$\inf_{T_\# f = g} \int_X c(x, T(x)) f(x)\, dx. \tag{6.60}$$

If c is defined by $c(x,y) = |x-y|^2$, this problem reduces to (6.54). It is also equivalent to (6.54) when $c(x,y) = -\langle x, y\rangle$, as follows directly from the equality $|x-y|^2 = |x|^2 + |y|^2 - 2\langle x, y\rangle$.

Under suitable assumptions (see [DF14; MTW05]), there exists a solution $T\colon X \to \overline{Y}$ to (6.60) of the form $T(x) = c\text{-exp}_x(Du(x))$, where for any $x \in X$ and p, $y \in \mathbb{R}^d$, the function $c\text{-exp}_x\colon \mathbb{R}^d \to \mathbb{R}^d$ is such that

$$y = c\text{-exp}_x(p) \iff p = -D_x c(x,y), \tag{6.61}$$

and where the function u (called the *potential function*) is c-convex, in the sense that for any $x_0 \in X$, there exists $y_0 \in \mathbb{R}^d$ and $z_0 \in \mathbb{R}$ such that

$$u(x_0) = -c(x_0, y_0) - z_0, \qquad u(x) \geq -c(x, y_0) - z_0 \quad \text{in } X.$$

If $c(x,y) = -\langle x, y\rangle$, c-convexity coincides with the usual notion of convexity. In the general setting, if u is smooth enough then it may be shown to be a solution to the Monge-Ampère equation (6.1), with

$$A(x,p) = -D_{xx} c(x, c\text{-exp}_x(p)), \tag{6.62}$$

$$B(x,p) = \frac{f(x)}{g(c\text{-exp}_x(p))} |\det D_{xy} c(x, c\text{-exp}_x(p))|, \tag{6.63}$$

and the constraint that $T(x) = c\text{-exp}_x(Du(x)) \in \overline{Y}$, for any $x \in X$, may be written as (6.22), where for any $x \in \overline{X}$,

$$P(x) = -D_x c(x, Y). \tag{6.64}$$

Then a suitable choice of function φ in Proposition 6.3.6 would be $\varphi(x) := -D_x c(x, y_0)$ (or a mollification of it), for some $y_0 \in Y$.

6.5.2 Weak solutions to the Monge-Ampère equation

If the open set X is convex, and if $u\colon X \to \mathbb{R}$ is a convex function and E is a subset of X, then we denote by $\partial u(E)$ the union $\bigcup_{x \in E} \partial u(x)$, where $\partial u(x)$ is the subgradient of u at point x.

A notion of weak solutions to the Monge-Ampère equation that is directly related to the optimal transport problem (6.54) is the one of *Brenier solutions*.

Definition 6.5.2 (Brenier solution)**.** Assume (6.56), (6.58), and (6.59). A function $u\colon X \to \mathbb{R}$ is a *Brenier solution* to (6.1) and (6.22) if (i) it is convex and (ii) $(Du)_\# f = g$, in the sense that (6.55) holds for $T = Du$. It is a *minimal Brenier solution* if moreover $\partial u(X)$ is included in $\overline{Y}$.

Brenier solutions are a standard notion. Note that their definition allows that $Du(x) \notin \overline{Y}$, typically at points where $f(x) = 0$. Minimal Brenier solutions were introduced in [BD19] to prevent this and to guarantee uniqueness of solutions up to addition of a constant, as explained in the proof of [BD19, Proposition 3.1] (the proof uses the assumptions that Y is convex and g is nonnegative in Y):

Theorem 6.5.3 (Adapted from [BD19, Proposition 3.1])**.** *Assume* (6.56)*,* (6.58)*, and* (6.59)*. If u, $v\colon X \to \mathbb{R}$ are two minimal Brenier solutions to* (6.1) *and* (6.22)*, then there exists $\xi \in \mathbb{R}$ such that $v = u + \xi$.*

For any function $u\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, let us denote by $u^c\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ its Legendre-Fenchel transform, which we recall is defined by

$$u^c(y) := \sup_{x\in\mathbb{R}^d} (\langle x, y\rangle - u(x)).$$

If u is only defined in X (respectively $\overline{X}$), we define u^c in the same manner after having extended u with value $+\infty$ outside X (respectively $\overline{X}$). If $\tilde{Y}$ is a subset of $\mathbb{R}^d$, let us also define $u^{cc}_{\tilde{Y}}\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ by

$$u^{cc}_{\tilde{Y}}(x) := \sup_{y\in\tilde{Y}} (\langle x, y\rangle - u^c(y)),$$

so that $u^{cc} = u^{cc}_{\mathbb{R}^d}$. The motivation for the last definition is that under assumptions (6.56), (6.58), and (6.59), if $u\colon X \to \mathbb{R}$ is a Brenier solution to (6.1) and (6.22), then u^{cc}_Y is a minimal Brenier solution to (6.1) and (6.22).

Another standard notion of solutions to (6.1) and (6.22) is the one of *Aleksandrov solutions*:

Definition 6.5.4 (Aleksandrov solution)**.** Assume (6.56), (6.58), and (6.59). A function $u\colon X \to \mathbb{R}$ is an *Aleksandrov solution* to (6.1) and (6.22) if (i) it is convex and (ii) for any Borel subset E of X,

$$\int_E f(x)\,dx = \int_{Y\cap\partial u(E)} g(y)\,dy.$$

It is a *minimal Aleksandrov solution* to (6.1) and (6.22) if moreover $\partial u(X) \subset \overline{Y}$.

In our setting, Brenier and Aleksandrov solutions coincide, see for instance [FL09] (noting that the relevant part of [FL09] is not specific to the dimension two):

Proposition 6.5.5. *Assume* (6.56)*,* (6.58)*, and* (6.59)*. Then $u\colon X \to \mathbb{R}$ is a Brenier solution (respectively minimal Brenier solution) to* (6.1) *and* (6.22) *if and only if it is an Aleksandrov solution (respectively minimal Aleksandrov solution) to* (6.1) *and* (6.22)*.*

This is related to the fact that Y is convex and g is nonnegative in Y, and that this does not remain true in more general settings.

We will also need to use the notion of Aleksandrov solution to the Monge-Ampère equation equipped with the Dirichlet boundary condition

$$u(x) = \psi(x) \quad \text{on } \partial X. \tag{6.65}$$

Definition 6.5.6 (Aleksandrov solution to the Dirichlet problem)**.** Assume (6.56) and (6.58). A function $u\colon \overline{X} \to \mathbb{R}$ is an *Aleksandrov solution* to (6.1) and (6.65) if (i) it is convex continuous with $u(x) = \psi(x)$ on ∂X and (ii) for any Borel subset E of X,

$$\int_E f(x)\,dx = \int_{\partial u(E)} g(y)\,dy.$$

If $u\colon \overline{X} \to \mathbb{R}$ is continuous and is a minimal Aleksandrov solution to (6.1) and (6.22), then it is an Aleksandrov solution to (6.1) and (6.65) with $\psi = u_{|\partial X}$; however, this does not remain true if u is not minimal.

Below is the adaptation of [Gut16, Theorem 1.6.2] to our setting. For simplicity, it is assumed that $g(p) = 1$ for any $p \in \mathbb{R}^d$, but note that we only use Theorem 6.5.7 as an intermediary result and that our convergence result, Theorem 6.5.22, is not limited to the case $g(p) = 1$.

Theorem 6.5.7 (Adapted from [Gut16, Theorem 1.6.2])**.** *Assume* (6.58)*, that X is strictly convex, $g(p) = 1$ for any $p \in \mathbb{R}^d$, and $\psi\colon \partial X \to \mathbb{R}$ is continuous. Then there exists a unique Aleksandrov solution $u\colon \overline{X} \to \mathbb{R}$ to* (6.1) *and* (6.65)*.*

6.5.3 Reformulation of the Monge-Ampère equation

Let us now study the reformulation of the Monge-Ampère equation (6.1) in the form (6.2), in the setting of quadratic optimal transport. We sum up the idea of the reformulation in the following proposition:

Proposition 6.5.8. *Let $b \geq 0$ and $M \in \mathcal{S}_d^+$. Then*

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \leq 0 \iff b \leq \det M, \tag{6.66}$$

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \geq 0 \iff b \geq \det M. \tag{6.67}$$

Proof. We refer to [Kry87, Lemma 3.2.2] for the proof of the equivalence

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) = 0 \iff b = \det M. \tag{6.68}$$

Also, the first equality in (6.5) is proved in [Kry87, Lemma 3.2.1] (it is related to the inequality of arithmetic and geometric means applied to eigenvalues of the product $\mathcal{D}^{1/2}M\mathcal{D}^{1/2}$), while the second one follows from the identity

$$\{\mathcal{D} \in \mathcal{S}_d^{++} \mid \det \mathcal{D} = 1\} = \{(\det \mathcal{D})^{-1/d}\mathcal{D} \mid \mathcal{D} \in \mathcal{S}_d^{++},\, \mathrm{Tr}(\mathcal{D}) = 1\}.$$

From (6.5), we deduce that

$$\begin{aligned}
b \leq \det M &\iff db^{1/d} - d(\det M)^{1/d} \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} (db^{1/d} - (\det \mathcal{D})^{-1/d}\langle \mathcal{D}, M\rangle) \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} (db^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, M\rangle) \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \leq 0.
\end{aligned}$$

Then (6.66) follows from the continuity of $L_{\mathcal{D}}(b,M)$ with respect to $\mathcal{D} \in \mathcal{S}_d^+$, and (6.67) follows from (6.66) and (6.68). □

First we prove that Aleksandrov solutions to the Monge-Ampère equation are viscosity solutions to its reformulation.

Proposition 6.5.9. *Assume* (6.56) *and* (6.58). *If, for some function $\psi \in C(\partial X)$, $u\colon \overline{X} \to \mathbb{R}$ is an Aleksandrov solution to* (6.1) *and* (6.65), *then u is a viscosity solution to* (6.2).

The proof is an adaptation of the one of [Gut16, Proposition 1.3.4]. It uses [Gut16, Lemma 1.4.1], which we recall below in our setting:

Lemma 6.5.10. *Assume* (6.56). *Let u, $v\colon X \to \mathbb{R}$ be convex and let E be an open set such that $\overline{E} \subset X$. If $u \leq v$ in E and $u = v$ on ∂E, then $\partial v(E) \subset \partial u(E)$.*

Proof of Proposition 6.5.9. We adapt the proof of [Gut16, Proposition 1.3.4], which is a particular case of this proposition.

First let us show that u is a viscosity subsolution to (6.2). Let $\varphi \in C^2(X)$, and let $x_0 \in X$ be a local maximum of $u - \varphi$. Since u is convex, $D^2\varphi(x)$ must be positive semidefinite. We may assume without loss of generality that φ is convex, that $\varphi(x_0) = u(x_0)$, and that x_0 is a strict local maximum. For any small $\varepsilon > 0$, there exists an open set S_ε such that $\overline{S}_\varepsilon \subset X$, $\varphi \leq u + \varepsilon$ in S_ε, $\varphi = u + \varepsilon$ on ∂S_ε, and $\lim_{\varepsilon \to 0} d_H(S_\varepsilon, \{x_0\}) = 0$ (see [Gut16] for detail). By Lemma 6.5.10, $\partial u(S_\varepsilon) = \partial(u + \varepsilon)(S_\varepsilon) \subset \partial\varphi(S_\varepsilon)$. Thus, since u is an Aleksandrov solution,

$$\int_{S_\varepsilon} f(x)\,dx = \int_{\partial u(S_\varepsilon)} g(y)\,dy \leq \int_{\partial\varphi(S_\varepsilon)} g(y)\,dy = \int_{S_\varepsilon} g(D\varphi(x)) \det D^2\varphi(x)\,dx.$$

Passing to the limit in ε, we deduce that $f_*(x_0) \leq g(D\varphi(x_0)) \det D^2\varphi(x_0)$. By Proposition 6.5.8, it follows that $(F_{\mathrm{MA}})_*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \leq 0$, and thus that u is a viscosity subsolution to (6.2).

Now let us show that u is a viscosity supersolution to (6.2). Let $\varphi \in C^2(X)$, and let $x_0 \in X$ be a local minimum of $u - \varphi$. If there exists a unit vector $e \in \mathbb{R}^d$ such that $\langle e, D^2\varphi(x_0)e\rangle \leq 0$, then choosing $\mathcal{D} = e \otimes e$ in the maximum in the definition (6.8) of the operator F_{MA} yields

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq -\langle e, D^2\varphi(x_0)e\rangle \geq 0.$$

If on the contrary $D^2\varphi(x_0)$ is positive definite, then we may assume without loss of generality that φ is convex, that $\varphi(x_0) = u(x_0)$, and that x_0 is a strict local minimum. By the same reasoning as above, we prove that $f^*(x_0) \geq g(D\varphi(x_0)) \det D^2\varphi(x_0)$, and we deduce using Proposition 6.5.8 that $(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0$. Therefore u is a viscosity supersolution to (6.2). □

In order to prove convergence of a family of monotone numerical schemes for the Monge-Ampère equation, we need to study under which conditions viscosity subsolutions and supersolutions to (6.40) are minimal Aleksandrov solutions to (6.2) and (6.22). Thus the remaining part of this subsection is devoted to the proof of the two following theorems:

Theorem 6.5.11. *Assume* (6.56), (6.58), *and* (6.59). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity subsolution to* (6.40) *with* $\alpha \geq 0$, *then* $\alpha = 0$ *and* u *is a minimal Aleksandrov solution to* (6.1) *and* (6.22).

Theorem 6.5.12. *Assume* (6.57) *to* (6.59). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (6.40) *with* $\alpha \leq 0$, *then* $\alpha = 0$ *and* u_Y^{cc} *is a minimal Aleksandrov solution to* (6.1) *and* (6.22).

The result in the case of viscosity subsolutions is very close to [Fro19, Theorem 2.1] (although the reformulation of the Monge-Ampère equation is not the same) and thus we follow the same sketch of proof. The case of viscosity supersolutions was not studied in [Fro19] since it was not necessary for the proof of convergence of the scheme considered in that paper.

We will need the following comparison principle for equation (6.2):

Proposition 6.5.13 (Comparison principle)**.** *Assume that* $B^{1/d}$ *is continuous, in addition to being Lipschitz continuous with respect to its second variable, uniformly with respect to its first variable. Then there exists* $r > 0$ *such that the following holds: for any open subset* E *of* X *such that* $\operatorname{diam}(E) \leq r$ *and for any respectively upper and lower semicontinuous functions* $\overline{u}$, $\underline{u}\colon \overline{E} \to \mathbb{R}$, *if* $\overline{u}$ *and* $\underline{u}$ *are respectively a viscosity subsolution and a viscosity supersolution to*

$$F_{\mathrm{MA}}(x, Du(x), D^2u(x)) = 0 \quad \text{in } E,$$

and if $\overline{u} \leq \underline{u}$ *on* ∂E, *then* $\overline{u} \leq \underline{u}$ *in* E.

Proof. Let $x_0 \in E$. For any $\varepsilon > 0$, let $\underline{u}_\varepsilon \colon \overline{E} \to \mathbb{R}$ be defined by

$$\overline{u}_\varepsilon(x) := \overline{u}(x) + \frac{\varepsilon}{2}|x - x_0|^2 - \frac{\varepsilon}{2}\operatorname{diam}(E)^2,$$

so that $\overline{u}_\varepsilon \le \overline{u} \le \underline{u}$ on ∂E. Let $x_1 \in E$, $\varphi \in C^2(E)$, and $\varphi_\varepsilon := \varphi + (\varepsilon/2)|\cdot - x_0|^2$. Then x_1 is a local maximum of $\overline{u}_\varepsilon - \varphi_\varepsilon$ if and only if it is a local maximum of $\overline{u} - \varphi$. For some constant $C > 0$ and for $r = 1/(2C)$, using that $|D\varphi_\varepsilon(x_1) - D\varphi(x_1)| \le r\varepsilon$ and $D^2\varphi_\varepsilon(x_1) = D^2\varphi(x_1) + \varepsilon I_d$, it holds for any $\mathcal{D} \in \mathcal{S}_d^+$ satisfying $\operatorname{Tr}(\mathcal{D}) = 1$ that

$$\begin{aligned}
&L_{\mathcal{D}}\left(B(x, D\varphi_\varepsilon(x)), D^2\varphi_\varepsilon(x) - A(x, D\varphi_\varepsilon(x))\right)\\
&= dB(x, D\varphi_\varepsilon(x))^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, D^2\varphi_\varepsilon(x) - A(x, D\varphi_\varepsilon(x))\rangle\\
&\le dB(x, D\varphi(x))^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, D^2\varphi(x) - A(x, D\varphi(x))\rangle + Cr\varepsilon - \varepsilon\\
&= L_{\mathcal{D}}\left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) + Cr\varepsilon - \varepsilon\\
&\le L_{\mathcal{D}}\left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) - \varepsilon/2.
\end{aligned}$$

Thus if x_1 is a local maximum of $\overline{u}_\varepsilon - \varphi_\varepsilon$,

$$F_{\mathrm{MA}}(x_1, D\varphi_\varepsilon(x_1), D^2\varphi_\varepsilon(x_1)) \le F_{\mathrm{MA}}(x_1, D\varphi(x_1), D^2\varphi(x_1)) - \varepsilon/2 \le -\varepsilon/2.$$

Then by [CIL92, Theorem 3.3 and section 5.C], $\overline{u}_\varepsilon \le \underline{u}$ in E, and we conclude letting ε approach zero. □

Notice that we did not need to assume (6.58); however, if (6.58) holds, it may be shown that the assumption that $\operatorname{diam}(E) \le r$ is not necessary, see [IL90, Theorem V.2] for the argument.

We will also need the following lemmas.

Lemma 6.5.14. *Assume* (6.56) *and* (6.58). *If $u \colon X \to \mathbb{R}$ is a viscosity subsolution to* (6.2), *then it is convex.*

Proof. Let $\varphi \in C^2(X)$ and x_0 be a local maximum of $u - \varphi$ in X. Then, using that u is a viscosity subsolution and choosing $\mathcal{D} = e \otimes e$ in the maximum in the definition of F_{MA},

$$0 \ge (F_{\mathrm{MA}})_*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \ge -\min_{|e|=1}\langle e, D^2\varphi(x_0)e\rangle.$$

Thus u is a viscosity subsolution to

$$-\min_{|e|=1}\langle e, D^2u(x_0)e\rangle = 0 \quad \text{in } X.$$

By [Obe07, Theorem 1], it follows that u is convex. □

Lemma 6.5.15. *Assume* (6.56) *and* (6.59). *If $u \colon X \to \mathbb{R}$ is a convex viscosity subsolution to* (6.24), *then $\partial u(X) \subset \overline{Y}$.*

The proof of Lemma 6.5.15 is a direct transposition to our setting to the one of [Fro19, Lemma 2.5], so we do not reproduce it here.

Lemma 6.5.16. *Assume* (6.57), *i.e. that X is strongly convex. Then for any $x_0 \in \partial X$ and C, $\varepsilon > 0$, there exists a convex function $\psi \in C^2(\overline{X})$ such that x_0 is a local maximum of ψ and $|D\psi(x_0)| \le \varepsilon$, $\det D^2\psi(x_0) \ge C$.*

Proof. Since X is strongly convex, there exists $r > 0$ and a unit vector $e \in \mathbb{R}^d$, $|e| = 1$, such that $X \subset B_d(x_0 - re, r)$. Then for any $x \in \overline{X}$, one has $|x - (x_0 - re)|^2 \leq r^2$. Since $|x - (x_0 - re)|^2 = |x - x_0 + re|^2 = |x - x_0|^2 + 2r\langle e, x - x_0\rangle + r^2$, we deduce that $|x - x_0|^2 + 2r\langle e, x - x_0\rangle \leq 0$. Thus x_0 is a local maximum of $|\cdot - x_0|^2 + 2r\langle e, \cdot - x_0\rangle$ in $\overline{X}$. Therefore, using that $\langle e, \cdot - x_0\rangle \leq -|x - x_0|^2/(2r) < 0$ in X, x_0 is also a local maximum in $\overline{X}$ of the convex function $\varphi \in C^2(\overline{X})$ defined by

$$\begin{aligned}\psi(x) &:= \frac{\varepsilon}{4r}|x - x_0|^2 + \varepsilon\langle e, x - x_0\rangle + \frac{C}{2}\left(\frac{2r}{\varepsilon}\right)^{d-1}\langle e, x - x_0\rangle^2 \\ &= \frac{\varepsilon}{4r}\left(|x - x_0|^2 + 2r\langle e, x - x_0\rangle\right) + \frac{\varepsilon}{2}\langle e, x - x_0\rangle + \frac{C}{2}\left(\frac{2r}{\varepsilon}\right)^{d-1}\langle e, x - x_0\rangle^2.\end{aligned}$$

We compute that $|D\psi(x_0)| = \varepsilon$ and $\det D^2\psi(x_0) = (\varepsilon/(2r))^{d-1}(\varepsilon/(2r) + C(2r/\varepsilon)^{d-1}) \geq C$, which concludes the proof. □

Lemma 6.5.17. *Assume* (6.57) *to* (6.59). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (6.40) *with* $\alpha \leq 0$, *then* $Y \subset \partial u^{cc}(X)$.

Proof. Let $y_0 \in Y$. Since u is lower semicontinuous, there exists $x_0 \in \overline{X}$ such that $y_0 \in \partial u^{cc}(x_0)$ (meaning that x_0 is a local minimum of $u^{cc} - \langle \cdot, y_0\rangle$) and $u^{cc}(x_0) = u(x_0)$. Let us show that $x_0 \in X$.

Since $u^{cc} \leq u$ in $\overline{X}$, x_0 is a local minimum of $u - \langle \cdot, y_0\rangle$. If $x_0 \in \partial X$, then for any $\varepsilon > 0$, we may build using Lemma 6.5.16 a convex function $\varphi_\varepsilon \in C^2(\overline{X})$ such that x_0 is a local minimum of $u - \varphi_\varepsilon$ and

$$|D\varphi_\varepsilon(x_0) - y_0| \leq \varepsilon, \qquad \det D^2\varphi_\varepsilon(x_0) > \sup_{y \in \mathbb{R}^d} \frac{f^*(x_0)}{g(y)} \geq \frac{f^*(x_0)}{g(D\varphi_\varepsilon(x_0))}$$

(choose $\varphi_\varepsilon = \langle \cdot, y_0\rangle + \psi$ where ψ is from Lemma 6.5.16). Then by Proposition 6.5.8,

$$(F_{\mathrm{MA}})^*(x_0, D\varphi_\varepsilon(x_0), D^2\varphi_\varepsilon(x_0)) < 0.$$

We may choose ε small enough so that $D\varphi_\varepsilon(x_0) \in Y$, and thus $F_{\mathrm{BV2}}(x_0, D\varphi_\varepsilon(x_0)) < 0$. Then

$$(F^\alpha_{\mathrm{MABV2}})^*(x_0, D\varphi_\varepsilon(x_0), D^2\varphi_\varepsilon(x_0)) < 0,$$

which is impossible since u is a viscosity supersolution to (6.40). Therefore x_0 may not belong to ∂X. □

Lemma 6.5.18. *Assume* (6.56), (6.58), *and* (6.59). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (6.40) *with* $\alpha \leq 0$, *then* u^{cc}_Y *is a viscosity supersolution to* (6.2). *Moreover, if* $\alpha < 0$, $\varphi \in C^2(X)$, x_0 *is a local minimum of* $u^{cc}_Y - \varphi$ *in* X, *and* $f^*(x_0) > 0$, *then*

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) > 0.$$

Proof. Let $\varphi \in C^2(X)$, and let x_0 be a local minimum of $u^{cc}_Y - \varphi$ in X.

First we consider the case where $u^{cc}_Y(x_0) = u(x_0)$ and $\partial u^{cc}_Y(x_0) \subset Y$. Since $u^{cc}_Y \leq u$ in X, x_0 is a local minimum of $u - \varphi$ in X. Thus

$$(F^\alpha_{\mathrm{MABV2}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0.$$

Since $\partial u^{cc}_Y(x_0) \subset Y$, $D\varphi(x_0)$ belongs to Y. Therefore

$$F_{\mathrm{BV2}}(x_0, D\varphi(x_0)) < 0.$$

It follows that

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0,$$

with a strict inequality if $\alpha < 0$.

Now we consider the case where either $u_Y^{cc}(x_0) < u(x_0)$ or $\partial u_Y^{cc}(x_0) \cap \partial Y \neq \emptyset$. In this case, there exists a unit vector $e \in \mathbb{R}^d$ such that $\langle e, D^2\varphi(x)e\rangle \leq 0$. Choosing $\mathcal{D} = (1-\varepsilon)e \otimes e + (\varepsilon/d)I_d$ in the definition of F_{MA} yields

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq d\frac{f^*(x_0)^{1/d}}{g(D\varphi(x_0))^{1/d}}\left(1 - \frac{d-1}{d}\varepsilon\right)^{1/d}\varepsilon^{(d-1)/d} - \frac{\varepsilon}{d}\operatorname{Tr}(D^2\varphi(x_0)).$$

If $f^*(x_0) > 0$, we conclude by choosing ε small enough so that the right-hand side is positive. If $f^*(x_0) = 0$, we conclude by letting ε approach zero. □

Lemma 6.5.19. *Assume* (6.56) *and* (6.58). *If $u\colon X \to \mathbb{R}$ is a convex viscosity supersolution to* (6.2)*, then for any Borel subset E of X of Lebesgue measure zero, $\partial u(E)$ has Lebesgue measure zero.*

Proof. Let $K > 0$, and let E be a subset of X of Lebesgue measure zero. Then for any $\varepsilon > 0$, there exists an open set $G \subset X$ such that $E \subset G$ and $\mathcal{L}^d(G) \leq \varepsilon$. For any $x \in G$, let $r(x) > 0$ and $S(x) := B_d(x, r(x))$, choosing $r(x)$ small enough so that $\overline{S(x)} \subset G$. By Theorem 6.5.7, there exists an Aleksandrov solution $v \in C(\overline{S(x)})$ to

$$\begin{cases} \det{}_+ D^2 v(x) = K & \text{in } S(x), \\ v(x) = u(x) & \text{on } \partial S(x). \end{cases}$$

By Proposition 6.5.9, v is a viscosity solution to (6.2) with $A(x,p)$ replaced by zero, $B(x,p)$ replaced by K, and X replaced by E. Choosing K large enough, it is easily verified that u is a viscosity supersolution to (6.2) with the same parameters. Then by Proposition 6.5.13, up to choosing $r(x)$ smaller, $v \leq u$ in $S(x)$. Since $u = v$ on $\partial S(x)$, Lemma 6.5.10 shows that $\partial u(S(x)) \subset \partial v(S(x))$. Thus

$$\mathcal{L}^d(\partial u(S(x))) \leq \mathcal{L}^d(\partial v(S(x))) = K\mathcal{L}^d(S(x)).$$

Let $\delta < 1/5$ (for instance $\delta = 1/6$) and for any $x \in G$, let $S_\delta(x) \subset G$ be defined by $S_\delta(x) := B_d(x, \delta r(x))$. Then by Vitali's covering theorem [EG92, Theorem 1.5.1], there exists a countable family $(x_i)_{i\in\mathbb{N}}$ of points of G such that $\bigcup_{x\in G} S_\delta(x) \subset \bigcup_{i\in\mathbb{N}} S(x_i)$ and balls of the family $(S_\delta(x_i))_{i\in\mathbb{N}}$ are all disjoint. Since $E \subset G = \bigcup_{x\in G} S_\delta(x)$, we deduce that $\partial u(E) \subset \bigcup_{i\in\mathbb{N}} \partial u(S(x_i))$ and thus

$$\begin{aligned}\mathcal{L}^d(\partial u(E)) &\leq \sum_{i\in\mathbb{N}} \mathcal{L}^d(\partial u(S(x_i))) \leq K\sum_{i\in\mathbb{N}} \mathcal{L}^d(S(x_i)) = K\delta^{-d}\sum_{i\in\mathbb{N}} \mathcal{L}^d(S_\delta(x_i)) \leq K\delta^{-d}\mathcal{L}^d(G) \\ &\leq K\delta^{-d}\varepsilon.\end{aligned}$$

We conclude by letting ε approach zero that $\mathcal{L}^d(\partial u(E)) = 0$. □

Lemma 6.5.20. *Assume* (6.56). *If $u\colon X \to \mathbb{R}$ is convex, then the set*

$$\{y \in \mathbb{R}^d \mid \exists x_1, x_2 \in X,\ x_1 \neq x_2 \text{ and } y \in \partial u(x_1) \cap \partial u(x_2)\}$$

has Lebesgue measure zero.

Proof. This standard result follows directly from the facts that u^c is not twice differentiable at points of this set (since $\{x_1, x_2\} \subset \partial u^c(y)$) and that u^c, as a convex, hence locally Lipschitz function, is differentiable almost everywhere, by Rademacher's theorem [EG92, Theorem 3.1.2]. □

In the lemma below, the right-hand side in (6.69) is to be understood as the integral of function which coincides almost everywhere with $g(Du(\cdot)) \det D^2u(\cdot)$, the convex function u being twice differentiable almost everywhere by Aleksandrov's theorem [EG92, Theorem 6.4.1]. In particular, points where u is not twice differentiable do not contribute to the integral in the right-hand side, while they do contribute to the one in the left-hand side.

Lemma 6.5.21. *Assume* (6.56). *If $u\colon X \to \mathbb{R}$ is convex, then for any Borel subset E of X,*

$$\int_{\partial u(E)} g(y)\,dy \geq \int_E g(Du(x)) \det D^2u(x)\,dx. \tag{6.69}$$

If moreover $\partial u(E')$ has Lebesgue measure zero for any subset E' of X of Lebesgue measure zero, then the above inequality is an equality.

Proof. Since u is convex, its gradient Du belongs to $BV_{\mathrm{loc}}(X;\mathbb{R}^d)$, see [EG92, Theorem 6.3.3]. By [EG92, Theorem 6.6.2], for any $k \in \mathbb{N}^*$, there exists a subset E_k of E such that Du is Lipschitz continuous in E_k and $\mathcal{L}^d(E \setminus E_k) \leq 1/k$. We define $\tilde{E} := \bigcup_{k=1}^{\infty} E_k$ and, for any $k \in \mathbb{N}^*$, $\tilde{E}_k := E_k \setminus (\bigcup_{i=1}^{k-1} E_i)$.

Using Lemma 6.5.20,

$$\begin{aligned}\int_{\partial u(E)} g(y)\,dy \geq \int_{\partial u(\tilde{E})} g(y)\,dy &= \sum_{k=1}^{\infty} \int_{\partial u(\tilde{E}_k)} g(y)\,dy = \sum_{k=1}^{\infty} \int_{Du(\tilde{E}_k)} g(y)\,dy \\ &= \sum_{k=1}^{\infty} \int_{\mathbb{R}^d} \left[\sum_{x \in (Du)^{-1}(\{y\})} \mathbb{1}_{\tilde{E}_k(x)} g(Du(x)) \right] dy\end{aligned}$$

(here $(Du)^{-1}(\{y\})$ is a singleton for almost every y), with equality if $\partial u(E \setminus \tilde{E})$ has Lebesgue measure zero (note that $E \setminus \tilde{E}$ always has Lebesgue measure zero).

By the change of variables formula [EG92, Theorem 3.3.2], which is a corollary of the area formula of geometric measure theory, for any $k \in \mathbb{N}^*$,

$$\int_{\mathbb{R}^d} \left[\sum_{x \in (Du)^{-1}(\{y\})} \mathbb{1}_{\tilde{E}_k(x)} g(Du(x)) \right] dy = \int_{\tilde{E}_k} g(Du(x)) \det D^2u(x)\,dx.$$

It follows that

$$\int_{\partial u(\tilde{E})} g(y)\,dy = \sum_{k=1}^{\infty} \int_{\tilde{E}_k} g(Du(x)) \det D^2u(x)\,dx = \int_E g(Du(x)) \det D^2u(x)\,dx,$$

which concludes the proof. □

Let us now prove the main Theorem 6.5.11 and Theorem 6.5.12.

Proof of Theorem 6.5.11. If $u\colon \overline{X} \to \mathbb{R}$ is a viscosity subsolution to (6.40) with $\alpha \geq 0$, it is both a viscosity subsolution to (6.2) and (6.24). Thus by Lemma 6.5.14 and Lemma 6.5.15, it is convex in X and $\partial u(X) \subset \overline{Y}$.

By Aleksandrov's theorem [EG92, Theorem 6.4.1], u is twice differentiable almost everywhere. Thus it is almost everywhere a classical subsolution to (6.40). It follows that for almost every $x \in X$, $F_{\mathrm{MA}}(x, Du(x), D^2u(x)) \leq 0$, with a strict inequality if $\alpha > 0$. Then, using Proposition 6.5.8, for any Borel subset E of X,

$$\int_E f(x)\,dx \leq \int_E g(Du(x)) \det D^2u(x)\,dx,$$

with a strict inequality if $\alpha > 0$.

By Lemma 6.5.21,

$$\int_E f(x)\,dx \leq \int_{\partial u(E)} g(y)\,dy.$$

The same is true when replacing E by $X \setminus E$, and by Lemma 6.5.20, $\partial u(E) \cap \partial u(X \setminus E)$ has Lebesgue measure zero. But since $\partial u(X) \subset \overline{Y}$,

$$\int_{\partial u(X)} g(y)\,dy \leq \int_Y g(y)\,dy = \int_X f(x)\,dx.$$

It follows that

$$\int_E f(x)\,dx = \int_E g(Du(x)) \det D^2u(x)\,dx = \int_{\partial u(E)} g(y)\,dy.$$

Thus $\alpha = 0$ and u is a minimal Aleksandrov solution to (6.1) and (6.22). □

Proof of Theorem 6.5.12. When applicable, we follow the same sketch of proof as for Theorem 6.5.11. Let $u\colon \overline{X} \to \mathbb{R}$ be a viscosity supersolution to (6.40) with $\alpha \leq 0$. By Aleksandrov's theorem [EG92, Theorem 6.4.1], u_Y^{cc} is twice differentiable almost everywhere. Then by Lemma 6.5.18, for almost every $x \in X$, $F_{\mathrm{MA}}(x, Du_Y^{cc}(x), D^2u_Y^{cc}(x)) \geq 0$, with a strict inequality if $\alpha < 0$. Using Proposition 6.5.8, for any Borel subset E of X,

$$\int_E f(x)\,dx \geq \int_E g(Du_Y^{cc}(x)) \det D^2u_Y^{cc}(x)\,dx,$$

with a strict inequality if $\alpha < 0$ and $\mathcal{L}^d(\{x \in E \mid f(x) > 0\}) > 0$.

By Lemma 6.5.19 and Lemma 6.5.21,

$$\int_E f(x)\,dx \geq \int_{\partial u_Y^{cc}(E)} g(y)\,dy.$$

The same is true when replacing E by $X \setminus E$. But by Lemma 6.5.17, $Y \subset \partial u_Y^{cc}(X)$ and thus

$$\int_{\partial u_Y^{cc}(X)} g(y)\,dy = \int_Y g(y)\,dy = \int_X f(x)dx.$$

It follows that

$$\int_E f(x)\,dx = \int_E g(Du_Y^{cc}(x)) \det D^2u_Y^{cc}(x)\,dx = \int_{\partial u_Y^{cc}(E)} g(y)\,dy.$$

Thus $\alpha = 0$ and u_Y^{cc} is a minimal Aleksandrov solution to (6.1) and (6.22). □

6.5.4 Convergence

We are now able to prove convergence of a family of numerical schemes (which includes the scheme (6.28), see section 6.3) for the Monge-Ampère equation, in the setting of quadratic optimal transport.

Theorem 6.5.22 (Convergence)**.** *Assume* (6.57) *to* (6.59)*. If the scheme* (6.36) *is monotone, consistent with equation* (6.40)*, and equicontinuously stable (in the sense of Definition 6.2.12), and if for any small $h > 0$, there exists a solution $(\alpha_h, u_h) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to* (6.36) *satisfying $u_h[0] = 0$, then as h approaches zero, α_h converges to zero and u_h converges uniformly to the unique minimal Aleksandrov solution (or equivalently minimal Brenier solution) $u\colon X \to \mathbb{R}$ to* (6.1) *and* (6.22) *satisfying $u(0) = 0$.*

Proof. Let $(h_n)_{n\in\mathbb{N}}$ be a sequence of small discretization steps $h_n > 0$ converging to zero. Since (6.36) is equicontinuously stable, the sequence $(\alpha_{h_n})_{n\in\mathbb{N}}$ is bounded, and $(u_{h_n})_{n\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous. Then by the Arzelà-Ascoli theorem, up to extracting a subsequence, α_{h_n} converges to some $\alpha \in \mathbb{R}$ and u_{h_n} converges uniformly to some Lipschitz continuous function $u\colon \overline{X} \to \mathbb{R}$, satisfying $u(0) = 0$.

Let us show that $\alpha = 0$ and that u is a minimal Aleksandrov solution to (6.1) and (6.22). By Corollary 6.2.13, u is a viscosity solution (hence both a viscosity subsolution and supersolution) to (6.40). If $\alpha \leq 0$, then Theorem 6.5.12 implies that $\alpha = 0$. Thus we proved that $\alpha \geq 0$, and we may conclude by applying Theorem 6.5.11. □

6.6 Numerical application to nonimaging optics

We apply the finite difference scheme (6.28) to the far field refractor problem [GH09] in nonimaging optics. In this problem, and its variant, the near field refractor problem [GH14], light rays emanate from a light source located at the origin, in directions belonging to some subset $\hat{Y}$ of the unit sphere S^2, and with intensity described by a density $\hat{g}\colon \hat{Y} \to \mathbb{R}_+$. They propagate in an isotropic medium with index of refraction $n_1 > 0$, called medium I, until they hit a refractor, represented by a surface $\mathcal{R} \subset \mathbb{R}^3$. We will impose that $\mathcal{R}$ contains the point $e_3 = (0, 0, 1)$. The refracted rays then propagate in another isotropic medium, called medium II, with index of refraction $n_2 = \kappa n_1$, $0 < \kappa < 1$ (the case $\kappa > 1$, also studied in [GH09], has a different mathematical structure and is not addressed here). The refracted rays continue to propagate until they hit a screen, represented by the plane $\mathbb{R}^2 \times \{\ell\}$, for some $\ell > 0$. The aim is to find a suitable shape for the refractor $\mathcal{R}$ so that refracted rays hit the screen $\mathbb{R}^2 \times \{\ell\}$ at points belonging to $\ell(X \times \{1\})$, for some given subset X of $\mathbb{R}^2$, with intensity described by $\ell^{-2} f(\cdot/\ell)$, for some density $f\colon X \to \mathbb{R}_+$. Here we consider the far field problem, that is, the limit problem as ℓ approaches $+\infty$, while in the near field problem ℓ is a fixed finite number. We illustrate the problem in Figure 6.1.

Let us define $\psi\colon \mathbb{R}^2 \to S^2$ by

$$\psi(x) := \mathfrak{n}(x)(x, 1), \qquad \mathfrak{n}(x) := (1 + |x|^2)^{-1/2}, \tag{6.70}$$

so that $\psi(x)$ is the orthogonal projection of the point $(x, 1)$ onto the unit sphere (thus $\psi(x)$ is a unit vector, while $\mathfrak{n}(x)$ is a normalization factor). Then the far field refractor problem is equivalent to prescribing that light rays be refracted in directions belonging to the set $\hat{X} := \psi(X)$, with intensity described by a density $\hat{f}\colon \hat{X} \to \mathbb{R}_+$ such that

$$f(x) = J\psi(x)\hat{f}(\psi(x)) = \mathfrak{n}(x)^3 \hat{f}(\psi(x)),$$

where $J\psi$ is the Jacobian of ψ, in the sense of [EG92, section 3.2.2]. We will assume that there exists $Y \subset \mathbb{R}^2$ such that $\hat{Y} = \psi(Y)$, and define $g\colon Y \to \mathbb{R}_+$ by

$$g(y) := J\psi(y)\hat{g}(\psi(y)) = \mathfrak{n}(y)^3\hat{g}(\psi(y)). \tag{6.71}$$

It was shown in [GH09] that under suitable assumptions, including the mass balance condition (6.53) and the inequality

$$\inf_{\hat{x}\in\hat{X},\,\hat{y}\in\hat{Y}} \langle \hat{x}, \hat{y}\rangle \geq \kappa, \tag{6.72}$$

there exists an admissible refractor shape $\mathcal{R}$ to the far field refractor problem, of the form

$$\mathcal{R} = \{\exp(\kappa v(y))\psi(y) \mid y \in Y\}, \tag{6.73}$$

where $v\colon Y \to \mathbb{R}$ and $u\colon X \to \mathbb{R}$ are functions satisfying

$$v(y) = \sup_{x\in X}(-c(x,y) - u(x)), \qquad\qquad u(x) = \sup_{y\in Y}(-c(x,y) - v(y)), \tag{6.74}$$

with

$$c(x,y) := \frac{1}{\kappa}\log(1 - \kappa\langle\psi(x), \psi(y)\rangle), \tag{6.75}$$

and u is solution, in a generalized sense, to the Monge-Ampère equation (6.1), with the boundary condition (6.22) and with coefficients that we derive from [GH09] in section 6.C. The function v is solution to the same Monge-Ampère equation after reversing the roles of X and Y and of f and g, but numerically it is practical to discretize the equation satisfied u and not the one satisfied by v, since, in the setting considered below, the density g is Lipschitz continuous and uniformly nonzero on its domain, while f is not. As a remark, note that the solution to the near field refractor problem, that we do not approximate here, is described by the solution to a Monge-Ampère equation [GH14], but that this equation is of the form

$$\det{}_+\left(D^2u(x) - A(x, u(x), Du(x))\right) = B(x, u(x), Du(x)) \quad \text{in } X, \tag{6.76}$$

where in comparison with (6.1) the functions A and B feature an additional dependency with respect to u, and its set of solutions is stable by an invariance that is not the addition of a constant (this equation, as well as some other ones of the form (6.76), fit in the framework of *generated Jacobian equations* [Tru14]).

We consider the far field refractor problem with $\kappa = 2/3$, which is a typical value for a glass-air interface, source and target sets

$$\hat{Y} := \{\hat{y} \in S^2 \mid \hat{y}_3 > \delta_y\}, \qquad \hat{X} := \{\hat{x} \in S^2 \mid \hat{x}_3 > \delta_x\}, \qquad \delta_x = \delta_y = \cos(\pi/8),$$

corresponding to

$$Y = \{y \in \mathbb{R}^2 \mid |y|^2 < \delta_y^{-2} - 1\}, \qquad\qquad X = \{x \in \mathbb{R}^2 \mid |x|^2 < \delta_x^{-2} - 1\},$$

and with a uniform source density $\hat{g}(\hat{y}) = 1$, $\hat{y} \in \hat{Y}$, and a discontinuous target density f describing the image depicted in Figure 6.2, normalized so that (6.53) holds.

We approximate the pair $(0, u)$, where u is solution to (6.1) and (6.22) with the coefficients mentioned above, by a solution (α_h, u_h) to the numerical scheme (6.28) and to $u_h[0] = 0$ on the intersection $\mathcal{G}_h$ of the set X and of an $N \times N$ square Cartesian grid, where $N = 120$. More precisely, $\mathcal{G}_h := X \cap h\mathbb{Z}^d$, where

$$h := \frac{\operatorname{diam}(X)}{N} = \frac{2\sqrt{\delta_x^{-2} - 1}}{N} \approx 0.0069.$$

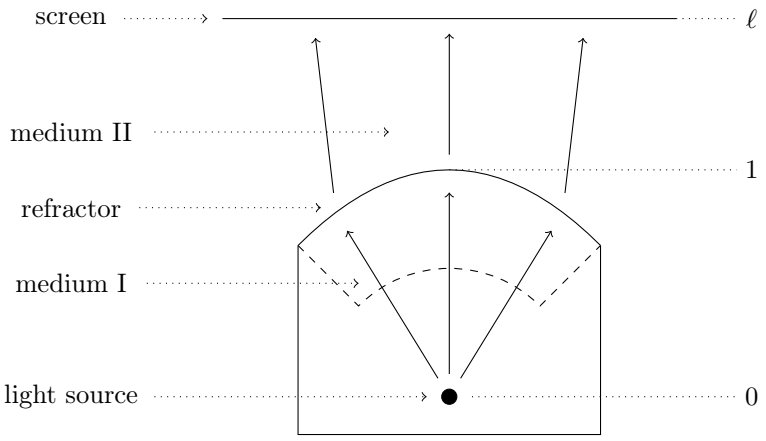


Figure 6.1: The far field refractor problem. Note that while in theory the light source belongs to medium I, in practice a second, spherical interface, represented by the dashed line, may be added between media I and II, since it would not refract light rays.

In (6.16) and (6.17), we choose $a_{\min} = -\infty$ and $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$. These parameters do not fit in the theoretical framework, but this does not seem to be a problem in practice for our application (recall that a_{LF} and b_{LF} are Lax-Friedrichs relaxation parameters and that choosing them as zero improves consistency of the scheme but fails to guarantee its monotonicity, see Proposition 6.3.1 and Remark 6.3.4; recall also that the finiteness of $a_{\min}$ is used in the proof of Proposition 6.3.6).

We let $\mu := 2 + \sqrt{5} \approx 4.24$, and we choose $V_h = V^\mu$, where the set V^μ, defined in section 6.B, contains the following superbases:

$$\left(\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right), \qquad \left(\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix},\begin{pmatrix}1\\1\end{pmatrix}\right),$$
$$\left(\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}-1\\2\end{pmatrix}\right), \qquad \left(\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}1\\2\end{pmatrix}\right),$$
$$\left(\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}-2\\1\end{pmatrix}\right), \qquad \left(\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix},\begin{pmatrix}2\\1\end{pmatrix}\right).$$

We compute the second maximum in (6.21) using Theorem 6.1.2. We discretize the maximum in (6.26) over the set $\{(\cos(k\pi/50), \sin(k\pi/50) \mid k \in \{-50, \dots, 50\}\}$; however, for computing $\sigma_{P(x)}(e)$, which is itself defined as a supremum, we use a closed-form formula, see section 6.C.

We use a Newton method to solve the scheme (6.28), together with the additional constraint $u[0] = 0$. More precisely, we look for a zero of the function $(\alpha, u) \mapsto S_h^\alpha u[x]$ over the hyperplane $\mathbb{R} \times \{u \in \mathbb{R}^{\mathcal{G}_h} \mid u[0] = 0\}$. We use the initialization

$$u_h^{(0)}(x) := -c(x, 0) = -\frac{1}{\kappa} \log(1 - \kappa \mathfrak{n}(x)),$$

which describes a refractor with the uniform refraction property, see [GH09]. The Newton method converges in 12 iterations, with the stopping criterion

$$\max_{x \in \mathcal{G}_h} |S_h^\alpha u[x]| < 10^{-11}.$$

Let us now explain how we approximate the refractor $\mathcal{R}$ itself. Formally, if x is optimal in the first supremum in (6.74), then $-D_x c(x, y) - Du(x) = 0$, which we rewrite as $y = c\text{-}\exp_x(Du(x))$,

Figure 6.2: Left: target image. Right: simulation of the scene using the appleseed® rendering engine; the small black disk at the bottom represents the light source.

using the notation introduced in (6.61). This yields the formula

$$v(c\text{-}\exp_x(Du(x))) = -c(x, c\text{-}\exp_x(Du(x))) - u(x).$$

This motivates us to approximate the graph of the function v by the set

$$\{(y_h(x), v_h(x)) \mid x \in \tilde{\mathcal{G}}_h\},$$

where

$$\tilde{\mathcal{G}}_h := \{x \in \mathcal{G}_h \mid x + he_i \in \mathcal{G}_h \text{ and } x - he_i \in \mathcal{G}_h,\ \forall i \in \{1,2\}\},$$

$$y_h(x) := c\text{-}\exp_x(D_h u_h(x)), \qquad v_h(x) := -c(x, c\text{-}\exp_x(D_h u_h(x))) - u_h(x),$$

and the operator D_h is defined in (6.12). We then define the set

$$\mathcal{R}_h := \{\exp(\kappa v_h(x))\psi(y_h(x)) \mid x \in \tilde{\mathcal{G}}_h\},$$

up to adding a constant to the function v_h so that $\mathcal{R}_h$ contains a point close to $e_3 := (0,0,1)$, and we approximate the refractor $\mathcal{R}$ by the graph of a function

$$\tilde{v}\colon \operatorname{Conv}(\{(\hat{y}_1, \hat{y}_2) \mid \hat{y} \in \mathcal{R}_h\}) \to \mathbb{R}$$

which is a cubic (Clough-Tocher) interpolation of the points of $\mathcal{R}_h$.

In order to validate the numerical results, we simulated the scene of the far field refractor problem using the appleseed®[1] rendering engine. The chosen refractor is a triangle mesh finely approximating the graph of $\tilde{v}$, and the screen is at distance $\ell = 10$ from the light source at the origin. We present the result of the simulation in Figure 6.2. The bright circle around the reconstructed image corresponds to light rays near the boundary that do not hit the refractor.

In Figure 6.3, we display the graph of the function $\tilde{v}$, as well as a finite difference approximation of its pointwise Gaussian curvature, defined formally by the map

$$\tilde{y} \mapsto \frac{\det D^2\tilde{v}(\tilde{y})}{(1 + |D\tilde{v}(\tilde{y})|^2)^2}.$$

[1] `https://appleseedhq.net/`

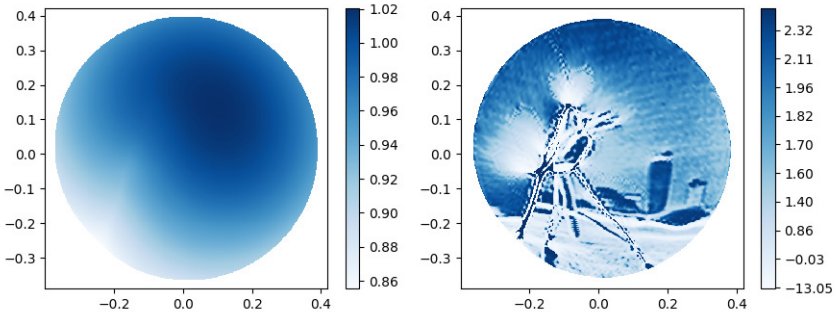


Figure 6.3: Approximation of the refractor (left) and of its pointwise curvature (right). Note that in the case of the curvature, the scale on the right is not linear.

Here the finite differences do not need to be monotone, thus we simply approximate separately all elements of the Hessian of $\tilde{v}$, and we use the standard formula for computing the determinant. Since $\tilde{v}$ is not expected to be twice differentiable, the finite difference approximation is not necessarily convergent, but it is informative nevertheless. We observe that the parts of the refractor corresponding to dark areas of the image have a small area, compared to the ones corresponding to bright areas. This is consistent with the fact that the total intensity of the light traversing them should be low, in order for the image to be properly reconstructed.

6.7 Conclusion and perspectives

We were able to adapt Perron's method in order to prove the existence of solutions to a class of monotone numerical schemes whose sets of solutions are stable by addition of a constant. We designed a finite difference scheme for the Monge-Ampère equation that belongs to this class, and proved convergence of the scheme in the setting of quadratic optimal transport. We showed that in dimension two, the discretization of the Monge-Ampère operator admits a closed-form formulation, and thus yields a particularly efficient numerical method, when carefully choosing its parameters using Selling's formula. We validated the method by numerical experiments in the context of the far field refractor problem in nonimaging optics.

A natural perspective is the adaptation of the proof of convergence of the scheme to the setting of more general optimal transport problems. The extension of the scheme to generated Jacobian equations such as (6.76) could also be studied. This would require adapting both the proof of convergence and the one of existence of solutions to the scheme, since the invariance in the set of solutions would not be the same in this case. Another perspective is studying how parameters of the discretization may be chosen to make the evaluation of the scheme efficient in dimensions higher than two, possibly using Selling's formula in dimension three or its counterpart, Voronoi's first reduction of quadratic forms [CS88], in dimensions four and higher.

6.A Relation to the MA-LBR scheme

MA-LBR is a numerical scheme for the two-dimensional Monge-Ampère equation, which was introduced in [BCM16]. Its natural extension to our setting would amount to replacing the

definition (6.21) of the operator $S^h_{\mathrm{MA}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ by

$$S^h_{\mathrm{MA}} u[x] := S^h_{\mathrm{adm}} u[x] \vee (B_h u[x] - \min_{v \in V_h} G(\Delta^v_h u[x] - A^v_h u[x])),$$

where V_h is a finite set of superbases of $\mathbb{Z}^2$, $S^h_{\mathrm{adm}} u \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ enforces a discrete version of the admissibility constraint (6.4), and the function $G \colon (\mathbb{R} \cup \{+\infty\})^3 \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$G(m) := \begin{cases} m_2 m_3 & \text{if } m_1 \geq m_2 + m_3, \\ m_1 m_3 & \text{if } m_2 \geq m_1 + m_3, \\ m_1 m_2 & \text{if } m_3 \geq m_1 + m_2, \\ \frac{1}{2}(m_1 m_2 + m_1 m_3 + m_2 m_3) & \\ \qquad -\frac{1}{4}(m_1^2 + m_2^2 + m_3^2) & \text{else.} \end{cases}$$

The operator S^h_{adm} is typically chosen as

$$S^h_{\mathrm{adm}} u[x] := - \min_{v \in E_h} (\Delta^e_h u[x] - A^e_h u[x]),$$

where E_h is a finite subset of $\mathbb{Z}^2$. Then any solution u to $S^h_{\mathrm{MA}} u[x] = 0$ in $\mathcal{G}_h$ satisfies $\Delta^e_h u[x] - A^e_h u[x] \geq 0$ in $\mathcal{G}_h$, for any $e \in E_h$.

Recall that for any $v \in (\mathbb{Z}^2)^3$, $\gamma \in \mathbb{R}^3$, and sufficiently smooth function u,

$$\langle \mathcal{D}_v(\gamma), D^2 u(x) \rangle = \sum_{i=1}^3 \gamma_i \langle v_i, D^2 u(x) v_i \rangle \approx \sum_{i=1}^3 \gamma_i \Delta^{v_i}_h u[x] = \langle \gamma, \Delta^v_h u[x] \rangle.$$

Thus the following proposition shows that the MA-LBR scheme may be seen as a discretization of the reformulation (6.6) of the Monge-Ampère equation:

Proposition 6.A.1. *If v is a superbase of $\mathbb{Z}^2$ and $m \in \overline{\mathbb{R}}^3_+$, then*

$$G(m) = \inf_{\substack{\gamma \in \mathbb{R}^3_+ \\ \det \mathcal{D}_v(\gamma) = 1}} \frac{\langle \gamma, m \rangle^2}{4}.$$

Proof. Using that v is a superbase of $\mathbb{Z}^2$, and thus for any $1 \leq i < j \leq 3$, $\det(v_i, v_j) = \pm 1$, we compute that for any $\gamma \in \mathbb{R}^3_+$,

$$\begin{aligned} \det \mathcal{D}_v(\gamma) &= \left(\sum_{i=1}^3 \gamma_i v_{i,1}^2 \right) \left(\sum_{i=1}^3 \gamma_i v_{i,2}^2 \right) - \left(\sum_{i=1}^3 \gamma_i v_{i,1} v_{i,2} \right)^2 \\ &= \sum_{i=1}^3 \sum_{j=1}^3 \gamma_i \gamma_j v_{i,1}^2 v_{j,2}^2 - \sum_{i=1}^3 \sum_{j=1}^3 \gamma_i \gamma_j v_{i,1} v_{i,2} v_{j,1} v_{j,2} \\ &= \sum_{i=1}^3 \sum_{j=1}^3 \gamma_i \gamma_j v_{i,1} v_{j,2} \det(v_i, v_j) \\ &= \sum_{1 \leq i < j \leq 3} \gamma_i \gamma_j v_{i,1} v_{j,2} \det(v_i, v_j) + \sum_{1 \leq i < j \leq 3} \gamma_j \gamma_i v_{j,1} v_{i,2} \det(v_j, v_i) \\ &= \sum_{1 \leq i < j \leq 3} \gamma_i \gamma_j \det(v_i, v_j)^2 \end{aligned}$$

$$= \sum_{1\le i<j\le 3} \gamma_i\gamma_j.$$

We conclude by noticing that

$$\inf_{\substack{\gamma\in\mathbb{R}^3_+ \\ \gamma_1\gamma_2+\gamma_1\gamma_3+\gamma_2\gamma_3=1}} \langle \gamma, m\rangle = 2G(m)^{1/2},$$

which is easily proved. □

6.B Choosing the set of superbases in dimension two

In this appendix, we explain how one may choose, in dimension $d = 2$ and for any $h > 0$, a finite set V_h of superbases of $\mathbb{Z}^2$ satisfying (6.18) to (6.20). The motivation is to use this set V_h in (6.21). The construction of V_h is based on the Stern-Brocot tree of bases of $\mathbb{Z}^2$ (see [BOZ04] for a similar approach in the setting of Hamilton-Jacobi-Bellman equations):

Definition 6.B.1. A pair (u, v) of vectors of $\mathbb{Z}^2$ is a *direct basis* of $\mathbb{Z}^2$ if $\det(u, v) = 1$.

Definition 6.B.2. The *Stern-Brocot tree* $\mathcal{T}$ is the collection of direct bases of $\mathbb{Z}^2$ defined inductively as follows: (i) the canonical basis belongs to $\mathcal{T}$, and (ii) for any $(u, v) \in \mathcal{T}$, one has $(u, u + v) \in \mathcal{T}$ and $(u + v, v) \in \mathcal{T}$.

Remark 6.B.3. In classical descriptions of the Stern-Brocot tree, the vector $u = (p, q)$ is often identified with the ratio p/q, which is a non-negative rational, or with $+\infty$, and likewise for $v = (r, s)$ (note that p and q are nonnegative and coprime by construction).

For any $(u, v) \in \mathcal{T}$, the scalar product $\langle u, v\rangle$ is a non-negative integer, as follows from an immediate induction. The set $\mathcal{T}_s := \{(u, v) \in \mathcal{T}; \langle u, v\rangle < s\}$ is a finite subtree which can be generated by exploration with the obvious stopping criterion, since $\min\{\langle u, u + v\rangle, \langle u + v, v\rangle\} = \langle u, v\rangle + \min\{|u^2|, |v|^2\} \ge \langle u, v\rangle + 1$.

Lemma 6.B.4. *Let $\mu > 1$ and $(u, v) \in \mathcal{T}_{(\mu-\mu^{-1})/2}$. Then*

$$\max\{|u|, |v|\} < \frac{\mu + \mu^{-1}}{2} < \mu.$$

Proof. It holds that

$$|u|^2 \le |u|^2|v|^2 = \det(u, v)^2 + \langle u, v\rangle^2 < 1 + \left(\frac{\mu - \mu^{-1}}{2}\right)^2 = \frac{\mu^2 + \mu^{-2} + 2}{4} = \left(\frac{\mu + \mu^{-1}}{2}\right)^2,$$

and similarly for v. □

For any $\mathcal{D} \in \mathcal{S}_2^{++}$, we define

$$\mu(\mathcal{D}) := \sqrt{|\mathcal{D}||\mathcal{D}^{-1}|}, \qquad s(\mathcal{D}) := \frac{1}{2}(\mu(\mathcal{D}) - \mu(\mathcal{D})^{-1}).$$

Note that $\mu(\mathcal{D})$ is the square root of the condition number of $\mathcal{D}$.

Lemma 6.B.5. *Let $(u, v) \in \mathcal{T}$ and $\mathcal{D} \in \mathcal{S}_2^{++}$. If $\langle u, v\rangle \ge s(\mathcal{D})$, then $\langle u, \mathcal{D}v\rangle \ge 0$.*

Proof. Denote by $\sphericalangle(u,v) \in [0,\pi]$ the unoriented angle between two vectors, defined by

$$\cos \sphericalangle(u,v) := \frac{\langle u,v\rangle}{|u||v|}.$$

On the one hand one has

$$\sin \sphericalangle(u,v) = \frac{\det(u,v)}{\sqrt{\langle u,v\rangle^2 + \det(u,v)^2}} = (1+\langle u,v\rangle^2)^{-1/2}.$$

On the other hand on can show [Des+21, Corollary B.4] that for any vector v,

$$(\mu(\mathcal{D}) + \mu(\mathcal{D})^{-1}) \cos \sphericalangle(v,\mathcal{D}v) \geq 2.$$

If $\langle u,v\rangle \geq (\mu(\mathcal{D})-\mu(\mathcal{D})^{-1})/2$, then one obtains $\sin\sphericalangle(u,v) \leq \cos\sphericalangle(v,\mathcal{D}v)$, and therefore $\sphericalangle(u,v)+\sphericalangle(v,\mathcal{D}v) \leq \pi/2$. By subadditivity of angles, $\sphericalangle(u,\mathcal{D}v) \leq \pi/2$, which is the announced result. □

Definition 6.B.6. Let $\mathcal{D} \in \mathcal{S}_2^+$. A superbase $v = (v_1,v_2,v_3)$ of $\mathbb{Z}^2$ is $\mathcal{D}$-*obtuse* if $\langle v_i, \mathcal{D}v_j\rangle \leq 0$, for any $1 \leq i < j \leq 3$.

Corollary 6.B.7. *For any $\mathcal{D} \in \mathcal{S}_2^{++}$, there exists $(u,v) \in \mathcal{T}$ such that $\langle u,v\rangle \leq s(\mathcal{D})$ and, denoting $\tilde{u} := (u_1,-u_2)$ and $\tilde{v} := (v_1,-v_2)$, either $(u,v,-u-v)$ or $(\tilde{u},\tilde{v},-\tilde{u}-\tilde{v})$ is a $\mathcal{D}$-obtuse superbase.*

Proof. We can assume that the non-diagonal coefficient of $\mathcal{D}$ is negative, up to reversing the orientation of one axis, and removing the trivial case of diagonal matrices. Let $(u,v) \in \mathcal{T}$ be such that $\langle u,\mathcal{D}v\rangle < 0$ and $\langle u,v\rangle$ is maximal. Such an element exists since the canonical basis obeys the condition $\langle u,\mathcal{D}v\rangle < 0$, since $\langle u,v\rangle$ is a non-negative integer, and since $\langle u,\mathcal{D}v\rangle \geq 0$ when $\langle u,v\rangle \geq s(\mathcal{D})$, by Lemma 6.B.5. Then, by construction, $\langle u,\mathcal{D}(u+v)\rangle \geq 0$ and $\langle u+v,\mathcal{D}v\rangle \geq 0$, which shows that $(u,v,-u-v)$ is a $\mathcal{D}$-obtuse superbase. □

For any $\mu > 1$, we define

$$V^\mu := \bigcup_{(u,v)\in\mathcal{T}_{(\mu-\mu^{-1})/2}} \{(-u^\perp,-v^\perp,u^\perp+v^\perp),(-\tilde{u}^\perp,-\tilde{v}^\perp,\tilde{u}^\perp+\tilde{v}^\perp)\},$$

where $\tilde{u} := (u_1,-u_2)$ and $\tilde{v} := (v_1,-v_2)$. The construction of the set V^μ is motivated by the following observation: if $\mathcal{D} \in \mathcal{S}_d^{++}$ obeys $\mu(\mathcal{D}) < \mu$, then, using Corollary 6.B.7 and that $s(\mathcal{D}) < (\mu-\mu^{-1})/2$, there exists a superbase $v = (v_1,v_2,v_3) \in V^\mu$ such that $(v_1^\perp,v_2^\perp,v_3^\perp)$ is $\mathcal{D}$-obtuse.

One may choose a sequence $(\mu_h)_{h>0}$ of parameters $\mu_h > 1$, and let $V_h = V^{\mu_h}$.

Proposition 6.B.8. *For any $h > 0$, let $\mu_h > 1$ be such that*

$$\lim_{h\to 0} \mu_h = +\infty, \qquad\qquad \lim_{h\to 0} h\mu_h = 0,$$

and let $V_h = V^{\mu_h}$. Then (6.18) to (6.20) are satisfied.

Proof. For fixed $h > 0$, let $\mathcal{D} \in \mathcal{S}_2^{++}$ be such that $\mu(\mathcal{D}) < \mu_h$. Then there exists a superbase $v = (v_1,v_2,v_3) \in V_h = V^{\mu_h}$ such that $(v_1^\perp,v_2^\perp,v_3^\perp)$ is $\mathcal{D}$-obtuse. By Selling's formula Proposition 6.4.2, there exists $\gamma \in \mathbb{R}_+^3$ such that $\mathcal{D} = \mathcal{D}_v(\gamma)$ (choose $\gamma = \gamma_v(\mathcal{D})$). It follows that

$$\{\mathcal{D} \in \mathcal{S}_2^{++} \mid \mathrm{Tr}(\mathcal{D}) = 1,\ \mu(\mathcal{D}) < \mu_h\} \subset \{\mathcal{D}_v(\gamma) \mid v \in V_h,\ \gamma \in \mathbb{R}_+^3,\ \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}.$$

Therefore

$$\begin{aligned}&\lim_{h\to 0} d_H(\{\mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}^3_+,\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}, \{\mathcal{D} \in \mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\})\\ &\leq \lim_{h\to 0} d_H(\{\mathcal{D} \in \mathcal{S}_2^{++} \mid \mathrm{Tr}(\mathcal{D}) = 1,\, \mu(\mathcal{D}) \leq \mu_h\}, \{\mathcal{D} \in \mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\})\\ &= 0,\end{aligned}$$

which proves (6.18).

Let $v = (v_1, v_2, v_3)$ be a superbase belonging to V_h. By Lemma 6.B.4, $\max_{1\leq i\leq 3} |v_i| \leq 2\mu_h$, and (6.19) follows.

Finally, (6.20) is satisfied since the subtree $\mathcal{T}_{(\mu_h - \mu_h^{-1})/2}$ always contains the canonical basis (e_1, e_2), hence $(-e_2, e_1, e_2 - e_1) = (-e_1^\perp, -e_2^\perp, e_1^\perp + e_2^\perp) \in V_h$. □

Remark 6.B.9. Let $c > 0$, $r \in (0, 1)$, and, for sufficiently small $h > 0$, choose $V_h = V^{\mu_h}$ where $\mu_h := ch^{-r}$. Then the proof of Proposition 6.B.8 yields the following refinements of (6.18) and (6.19):

$$\begin{aligned}&d_H\left(\{\mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}^3_+,\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}, \{\mathcal{D} \in \mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\}\right) = O(h^{2r}),\\ &\max_{v\in V_h} \max_{e\in v} |e| = O(h^{-r}),\end{aligned}$$

where the exponent in the first formula may be obtained by rewriting the relevant part of (6.52) as $1 - |\rho| = 2/(\mathrm{Cond}(\mathfrak{D}(\rho)) - 1) = 2/(\mu(\mathfrak{D}(\rho))^2 - 1) = O(\mu(\mathfrak{D}(\rho)))^{-2}$.

Let us give the following upper bound on the cardinal of the set V^μ:

Proposition 6.B.10. *There exists $C > 0$ such that for any $\mu > 1$, one has $\#(V^\mu) \leq C\mu(1 + \log \mu)$.*

Proof. By [Mir14, Lemma 2.7], there exists $C > 0$ such that for any $s > 1$, one has $\#(\mathcal{T}_s) \leq Cs(1 + \log s)$. The stated result follows, since $\#(V^\mu) = 2\#(\mathcal{T}_{(\mu-\mu^{-1})/2})$ and $\mathcal{T}_{(\mu-\mu^{-1})/2} \subset \mathcal{T}_\mu$. □

6.C Coefficients of the Monge-Ampère equation in the far field refractor problem

In this appendix, we compute the coefficients of the Monge-Ampère equation associated to the far field refractor problem that we solve numerically in section 6.6. It was shown in [GH09] that under suitable assumptions, the far field refractor problem admits a solution of the form (6.73) and (6.74), where u is the potential function of the optimal transport problem (6.60) with a cost function c defined by (6.75) in the domain $\overline{X} \times \overline{Y}$. This means that u is solution, in a generalized sense, to the Monge-Ampère equation (6.1), with the boundary condition (6.22) and with coefficients defined by (6.62) to (6.64).

Recall that $X, Y \subset \mathbb{R}^2$, $\hat{X}, \hat{Y} \subset S^2$, and $\hat{X} = \psi(X)$, $\hat{Y} = \psi(Y)$, where ψ is defined in (6.70). Let us denote by $\hat{c}$ the real-valued function defined in a neighborhood of $\hat{X} \times \hat{Y}$ by

$$\hat{c}(\hat{x}, \hat{y}) := \frac{1}{\kappa} \log\left(1 - \kappa \frac{\langle \hat{x}, \hat{y}\rangle}{|\hat{x}||\hat{y}|}\right).$$

Then, for any $x \in \overline{X}$ and $y \in \overline{Y}$,

$$c(x, y) = \hat{c}(\psi(x), \psi(y)). \tag{6.77}$$

In [GH09], the optimal transport problem was expressed on domains $\hat{X}$ and $\hat{Y}$ with the cost function $\hat{c}$, rather than on X and Y with the cost function c. We consider the problem on X and Y, since those domains may be discretized as two-dimensional Cartesian grids while $\hat{X}$ and $\hat{Y}$ may not. This requires adapting to our setting the formulae that were obtained in [GH09] for the coefficients of the Monge-Ampère equation.

Proposition 6.C.1. *Assume* (6.72) *and let* $x \in \overline{X}$. *Then the map* $\overline{Y} \to -D_x c(x, \overline{Y})$, $y \mapsto -D_x c(x, y)$ *is a bijection. If* $c\text{-exp}_x \colon -D_x c(x, \overline{Y}) \to \overline{Y}$ *denotes its inverse bijection, and if* A *and* B *are functions defined respectively by* (6.62) *and* (6.63), *then for any* $p \in -D_x c(x, \overline{Y})$,

$$c\text{-exp}_x(p) = \frac{\lambda \mathfrak{n}(x) x + (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} p}{\lambda \mathfrak{n}(x) - (1 - \kappa\lambda)\mathfrak{n}(x)^{-1}\langle p, x\rangle}, \tag{6.78}$$

$$A(x,p) = \kappa p \otimes p + \frac{\lambda}{1 - \kappa\lambda}(\mathfrak{n}(x)^4 x \otimes x - \mathfrak{n}(x)^2 I_2) - \mathfrak{n}(x)^2 (p \otimes x + x \otimes p), \tag{6.79}$$

$$B(x,p) = \frac{f(x)}{\mathfrak{m}\hat{g}(\hat{y})} \left| \frac{\mathfrak{n}(x)^4 - (\lambda + \kappa - \kappa\lambda^2)\mathfrak{n}(x)^2\langle p, x\rangle}{(1 - \kappa\lambda)^2} - \frac{|p|^2}{1 - \kappa\lambda} + \mathfrak{n}(x)^2 \det(p, x)^2 \right|, \tag{6.80}$$

where $\lambda = \lambda(x,p)$, $\mathfrak{m} = \mathfrak{m}(x,p)$, *and* $\hat{y} = \hat{y}(x,p)$ *are defined by*

$$\lambda = \frac{\kappa \mathfrak{n}(x)^{-2}(|p|^2 + \langle p, x\rangle^2) + \big(1 - (1 - \kappa^2)\mathfrak{n}(x)^{-2}(|p|^2 + \langle p, x\rangle^2)\big)^{1/2}}{1 + \kappa^2 \mathfrak{n}(x)^{-2}(|p|^2 + \langle p, x\rangle^2)}, \tag{6.81}$$

$$\mathfrak{m} = \lambda \mathfrak{n}(x) - (1 - \kappa\lambda)\mathfrak{n}(x)^{-1}\langle p, x\rangle, \tag{6.82}$$

$$\hat{y} = \lambda \mathfrak{n}(x) \begin{pmatrix} x \\ 1 \end{pmatrix} + (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} \begin{pmatrix} p \\ -\langle p, x\rangle \end{pmatrix}. \tag{6.83}$$

Proposition 6.C.2. *Assume* (6.72) *and that* Y *is the centered Euclidean ball of radius* $(\delta_y^{-2} - 1)^{1/2}$, *for some* $\delta_y \in (0,1)$. *Let* P *be the set-valued function defined by* (6.64). *Then for any* $x \in \overline{X}$ *and any* $e \in \mathbb{R}^2$ *of unit norm,*

$$\sigma_{P(x)}(e) = \frac{\mathfrak{n}(x)\delta_y\langle e, y_*\rangle - \mathfrak{n}(x)^3\delta_y\langle x, y_*\rangle\langle e, x\rangle - \mathfrak{n}(x)^3\delta_y\langle e, x\rangle}{1 - \kappa\mathfrak{n}(x)\delta_y\langle x, y_*\rangle - \kappa\mathfrak{n}(x)\delta_y}, \tag{6.84}$$

where $y_* = y_*(x,e)$ *and* $\mathfrak{f} = \mathfrak{f}(x,e)$ *are defined by*

$$y_* = \frac{((\delta_y^{-2} - 1)|\mathfrak{f}|^2 - \kappa^2 \mathfrak{n}(x)^4 (1 - \delta_y^2)^2 \det(e, x)^2)^{1/2}\mathfrak{f}}{|\mathfrak{f}|^2} + \frac{\kappa \mathfrak{n}(x)^2 (1 - \delta_y^2)\det(e, x)\mathfrak{f}^\perp}{|\mathfrak{f}|^2}, \tag{6.85}$$

$$\mathfrak{f} = (1 - \kappa\mathfrak{n}(x)\delta_y)\mathfrak{n}(x)\delta_y e - \mathfrak{n}(x)^3\delta_y\langle e, x\rangle x. \tag{6.86}$$

The motivation for Proposition 6.C.2 is that $\sigma_{P(x)}(e)$ is part of the definition (6.26) of the operator S^h_{BV2}, and that this is the only occurrence of the function P in the definition of the scheme.

The rest of this section is devoted to the proofs of Propositions 6.C.1 and 6.C.2. Those proofs are based on the chain rule for differentiating composite functions, some simplifications based on identities from linear algebra such as (6.105), and the study of a constrained optimization problem in Lemma 6.C.3. A natural objective, if the proposed numerical scheme is adapted to other settings in optimal transport or optics, is to automatize part of the construction of the coefficients of the Monge-Ampère equation by taking advantage of machine symbolic computation and automatic differentiation.

Proof of Proposition 6.C.1. Let $x \in \overline{X}$, $y \in \overline{Y}$, $p := -D_x c(x,y)$, $\hat{x} := \psi(x)$, $\hat{y} := \psi(y)$, and $\hat{p} := -D_{\hat{x}}\hat{c}(\hat{x},\hat{y})$.

By (6.77) and the chain rule, using implicit summation on repeated indices,

$$p = D\psi_i(x)\hat{p}_i. \tag{6.87}$$

On the other hand, it is easily verified from the definition of $\hat{c}$ that

$$\langle \hat{p}, \hat{x} \rangle = 0. \tag{6.88}$$

Therefore it holds that

$$\begin{pmatrix} p \\ 0 \end{pmatrix} = \begin{pmatrix} D\psi_i(x) \\ -\mathfrak{n}(x)^2 \hat{x}_i \end{pmatrix} \hat{p}_i = \mathfrak{n}(x) \begin{pmatrix} I_2 - \mathfrak{n}(x)^2 x \otimes x & -\mathfrak{n}(x)^2 x \\ -\mathfrak{n}(x)^2 x^\top & -\mathfrak{n}(x)^2 \end{pmatrix} \hat{p}.$$

Inverting this system yields

$$\hat{p} = \mathfrak{n}(x)^{-1} \begin{pmatrix} I_2 & -x \\ -x^\top & -1 \end{pmatrix} \begin{pmatrix} p \\ 0 \end{pmatrix} = \mathfrak{n}(x)^{-1} \begin{pmatrix} p \\ -\langle p, x \rangle \end{pmatrix}. \tag{6.89}$$

It was proved in [GH09] that

$$\hat{y} = \lambda \hat{x} + (1 - \kappa\lambda)\hat{p}, \tag{6.90}$$

$$\langle \hat{x}, \hat{y} \rangle = \lambda, \tag{6.91}$$

$$-D_{\hat{x}\hat{x}}\hat{c}(\hat{x},\hat{y}) = \kappa \left(\hat{p} + \frac{\lambda}{1-\kappa\lambda}\hat{x} \right) \otimes \left(\hat{p} + \frac{\lambda}{1-\kappa\lambda}\hat{x} \right) - \frac{\lambda}{1-\kappa\lambda} I_3, \tag{6.92}$$

$$-D_{\hat{x}\hat{y}}\hat{c}(\hat{x},\hat{y}) = \frac{I_3}{1-\kappa\lambda} + \frac{\kappa\hat{y}\otimes\hat{x} - \hat{x}\otimes\hat{x} - \hat{y}\otimes\hat{y} + \lambda\hat{x}\otimes\hat{y}}{(1-\kappa\lambda)^2}, \tag{6.93}$$

where

$$\lambda := \frac{\kappa|\hat{p}|^2 + \mathfrak{h}^{1/2}}{1 + \kappa^2|\hat{p}|^2}, \qquad \mathfrak{h} := 1 - (1-\kappa^2)|\hat{p}|^2. \tag{6.94}$$

Note that (6.91) follows directly from (6.88) and (6.90).

We deduce (6.81) from (6.89) and (6.94). We deduce (6.83) from the definition of $\hat{x}$ and from (6.89) and (6.90).

Since $\hat{y} = (\mathfrak{n}(y)y, \mathfrak{n}(y))$, it follows from (6.83) that

$$\mathfrak{n}(y)y = \lambda\mathfrak{n}(x)x + (1-\kappa\lambda)\mathfrak{n}(x)^{-1}p, \tag{6.95}$$

$$\mathfrak{n}(y) = \lambda\mathfrak{n}(x) - (1-\kappa\lambda)\mathfrak{n}(x)^{-1}\langle p, x \rangle. \tag{6.96}$$

Dividing (6.95) by (6.96), we deduce that $y = c\text{-}\exp_x(p)$ satisfies (6.78), and thus that it is uniquely determined by x and p.

We may rewrite (6.91) as

$$\mathfrak{n}(x)\mathfrak{n}(y)(\langle x, y \rangle + 1) = \lambda. \tag{6.97}$$

We compute that

$$D\psi(x) = \begin{pmatrix} \mathfrak{n}(x)I_2 - \mathfrak{n}(x)^3 x \otimes x \\ -\mathfrak{n}(x)^3 x^\top \end{pmatrix}, \qquad D\psi(y) = \begin{pmatrix} \mathfrak{n}(y)I_2 - \mathfrak{n}(y)^3 y \otimes y \\ -\mathfrak{n}(y)^3 y^\top \end{pmatrix}.$$

Therefore, using (6.97) for (6.100), (6.101), and (6.103),

$$D\psi_i(x)\hat{x}_i = \mathfrak{n}(x)^2 x - \mathfrak{n}(x)^4(|x|^2+1)x = 0, \tag{6.98}$$

$$D\psi_i(y)\hat{y}_i = \mathfrak{n}(y)^2 y - \mathfrak{n}(y)^4(|y|^2+1)y = 0, \tag{6.99}$$

$$\begin{aligned} D\psi_i(y)\hat{x}_i &= \mathfrak{n}(x)\mathfrak{n}(y)x - \mathfrak{n}(x)\mathfrak{n}(y)^3(\langle x,y\rangle+1)y \\ &= \mathfrak{n}(x)\mathfrak{n}(y)x - \lambda\mathfrak{n}(y)^2 y, \end{aligned} \tag{6.100}$$

$$\begin{aligned} D\psi_i(x)\hat{y}_i &= \mathfrak{n}(x)\mathfrak{n}(y)y - \mathfrak{n}(x)^3\mathfrak{n}(y)(\langle x,y\rangle+1)x \\ &= \mathfrak{n}(x)\mathfrak{n}(y)y - \lambda\mathfrak{n}(x)^2 x, \end{aligned} \tag{6.101}$$

$$\begin{aligned} D\psi_i(x)\otimes D\psi_i(x) &= \mathfrak{n}(x)^2 I_2 - 2\mathfrak{n}(x)^4 x\otimes x + \mathfrak{n}(x)^6(|x|^2+1)x\otimes x \\ &= \mathfrak{n}(x)^2 I_2 - \mathfrak{n}(x)^4 x\otimes x, \end{aligned} \tag{6.102}$$

$$\begin{aligned} D\psi_i(x)\otimes D\psi_i(y) &= \mathfrak{n}(x)\mathfrak{n}(y)I_2 - \mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y \\ &\quad + \mathfrak{n}(x)^3\mathfrak{n}(y)^3(\langle x,y\rangle+1)x\otimes y \\ &= \mathfrak{n}(x)\mathfrak{n}(y)I_2 - \mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y \\ &\quad + \lambda\mathfrak{n}(x)^2\mathfrak{n}(y)^2 x\otimes y. \end{aligned} \tag{6.103}$$

We also compute that

$$D_{x_ix_j}\psi(x) = \begin{pmatrix} 3\mathfrak{n}(x)^5 x_i x_j x - \mathfrak{n}(x)^3(\delta_{ij}x + x_i e_j + x_j e_i) \\ 3\mathfrak{n}(x)^5 x_i x_j - \mathfrak{n}(x)^3\delta_{ij} \end{pmatrix}.$$

Therefore, using (6.89),

$$\begin{aligned} D^2\psi_i(x)\hat{p}_i &= 3\mathfrak{n}(x)^4(\langle p,x\rangle - \langle p,x\rangle)x\otimes x - \mathfrak{n}(x)^2(\langle p,x\rangle - \langle p,x\rangle)I_2 \\ &\quad - \mathfrak{n}(x)^2(p\otimes x + x\otimes p) \\ &= -\mathfrak{n}(x)^2(p\otimes x + x\otimes p). \end{aligned} \tag{6.104}$$

Formula of $A(x,p)$. Using (6.62) for the first equality, (6.77) and the chain rule for the second one, ant the definition of $\hat{p}$ for the third one, we compute that

$$\begin{aligned} A(x,p) &= -D_{xx}c(x,y) \\ &= -D\psi_i(x)\otimes D\psi_j(x)D_{\hat{x}_i\hat{x}_j}\hat{c}(\hat{x},\hat{y}) - D^2\psi_i(x)D_{\hat{x}_i}\hat{c}(\hat{x},\hat{y}) \\ &= -D\psi_i(x)\otimes D\psi_j(x)D_{\hat{x}_i\hat{x}_j}\hat{c}(\hat{x},\hat{y}) + D^2\psi_i(x)\hat{p}_i. \end{aligned}$$

We deduce (6.79) using (6.87), (6.92), (6.98), (6.102), and (6.104).

Formula of $B(x,p)$. Using (6.77) and the chain rule for the first equality, (6.93), (6.98) to (6.101), and (6.103) for the second one, and (6.95) for the fourth one, we compute that

$$\begin{aligned} -D_{xy}c(x,y) &= -D\psi_i(x)\otimes D_{\psi_j}(y)D_{\hat{x}_i\hat{y}_j}\hat{c}(\hat{x},\hat{y}) \\ &= \frac{1}{1-\kappa\lambda}(\mathfrak{n}(x)\mathfrak{n}(y)I_2 - \mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y + \lambda\mathfrak{n}(x)^2\mathfrak{n}(y)^2 x\otimes y) \\ &\quad + \frac{\kappa}{(1-\kappa\lambda)^2}(\mathfrak{n}(x)^2\mathfrak{n}(y)^2 x\otimes y - \lambda\mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \lambda\mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y \\ &\qquad + \lambda^2\mathfrak{n}(x)^2\mathfrak{n}(y)^2 y\otimes x) \\ &= \frac{\mathfrak{n}(x)\mathfrak{n}(y)}{(1-\kappa\lambda)^2}((1-\kappa\lambda)I_2 - \mathfrak{n}(x)^2 x\otimes x - \mathfrak{n}(y)^2 y\otimes y + (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)\mathfrak{n}(y)x\otimes y \end{aligned}$$

$$
\begin{aligned}
&\quad + \kappa\lambda^2 \mathfrak{n}(x)\mathfrak{n}(y) y \otimes x) \\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y)}{1-\kappa\lambda} I_2 - \frac{\mathfrak{n}(x)^3\mathfrak{n}(y)}{1-\kappa\lambda} x \otimes x - \mathfrak{n}(x)^{-1}\mathfrak{n}(y) p \otimes p - \lambda \mathfrak{n}(x)\mathfrak{n}(y) p \otimes x \\
&\quad - \frac{\kappa \mathfrak{n}(x)\mathfrak{n}(y)}{1-\kappa\lambda} x \otimes p
\end{aligned}
$$

Using the formula

$$
\begin{aligned}
&\det(aI_2 - bx \otimes x - cp \otimes p - dp \otimes x - ex \otimes p) \\
&= a^2 - ab|x|^2 - ac|p|^2 - a(d+e)\langle p, x\rangle + (bc - de)\det(p, x)^2
\end{aligned}
\tag{6.105}
$$

with suitable coefficients a, b, c, d, $e \in \mathbb{R}$ for the first equality, and using that $\mathfrak{n}(x)^{-2} - |x|^2 = 1$ for the second one,

$$
\begin{aligned}
\det D_{xy} c(x,y) &= \frac{\mathfrak{n}(x)^2\mathfrak{n}(y)^2}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(x)^4\mathfrak{n}(y)^2|x|^2}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(y)^2|p|^2}{1-\kappa\lambda} - \frac{(\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\mathfrak{n}(y)^2\langle p, x\rangle}{(1-\kappa\lambda)^2} \\
&\quad + \mathfrak{n}(x)^2\mathfrak{n}(y)^2 \det(p,x)^2 \\
&= \frac{\mathfrak{n}(x)^4\mathfrak{n}(y)^2 - (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\mathfrak{n}(y)^2\langle p, x\rangle}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(y)^2|p|^2}{1-\kappa\lambda} \\
&\quad + \mathfrak{n}(x)^2\mathfrak{n}(y)^2 \det(p,x)^2.
\end{aligned}
$$

Thus, using (6.63) for the first equality and (6.71) for the second one,

$$
\begin{aligned}
B(x,p) &= \frac{f(x)}{g(y)} |\det D_{xy} c(x,y)| = \frac{f(x)}{\mathfrak{n}(y)^3 \hat{g}(\hat{y})} |\det D_{xy} c(x,y)| \\
&= \frac{f(x)}{\mathfrak{n}(y)\hat{g}(\hat{y})} \left| \frac{\mathfrak{n}(x)^4 - (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\langle p, x\rangle}{(1-\kappa\lambda)^2} - \frac{|p|^2}{1-\kappa\lambda} + \mathfrak{n}(x)^2 \det(p,x)^2 \right|.
\end{aligned}
$$

We define $\mathfrak{m} := \mathfrak{n}(y)$, so that the above immediately yields (6.80), and (6.82) is a simple rewriting of (6.96). □

In order to prove Proposition 6.C.2, we will need the following lemma.

Lemma 6.C.3. *Let $\alpha \in \mathbb{R}$, β, $\eta > 0$, $\mathfrak{a}$, $\mathfrak{b} \in \mathbb{R}^2$, and $\mathfrak{f} := \beta\mathfrak{a} - \alpha\mathfrak{b}$ be such that $\mathfrak{f} \neq 0$ and $\eta|\mathfrak{b}| < \beta$. Then*

$$
\operatorname*{argmax}_{|y|=\eta} \frac{\alpha + \langle \mathfrak{a}, y\rangle}{\beta + \langle \mathfrak{b}, y\rangle} = \frac{(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a},\mathfrak{b})^2)^{1/2}\mathfrak{f} - \eta^2 \det(\mathfrak{a},\mathfrak{b})\mathfrak{f}^\perp}{|\mathfrak{f}|^2}.
$$

Proof. Let $y_* \in \mathbb{R}^2$, $|y_*| = \eta$ belong to the argmax. Let us write the first order optimality condition:

$$
(\beta + \langle \mathfrak{b}, y_*\rangle)\mathfrak{a} - (\alpha + \langle \mathfrak{a}, y_*\rangle)\mathfrak{b} = \lambda y_*, \quad \lambda \in \mathbb{R}.
$$

Then

$$
\begin{aligned}
0 &= (\beta + \langle \mathfrak{b}, y_*\rangle)\langle \mathfrak{a}^\perp, y_*\rangle - (\alpha + \langle \mathfrak{a}, y_*\rangle)\langle \mathfrak{b}^\perp, y_*\rangle \\
&= \langle \beta\mathfrak{a}^\perp - \alpha\mathfrak{b}^\perp, y_*\rangle + \langle \mathfrak{b}, y_*\rangle\langle \mathfrak{a}^\perp, y_*\rangle - \langle \mathfrak{a}, y_*\rangle\langle \mathfrak{b}^\perp, y_*\rangle \\
&= \langle \mathfrak{f}^\perp, y_*\rangle + \det(\mathfrak{a},\mathfrak{b})|y_*|^2 = \langle \mathfrak{f}^\perp, y_*\rangle + \eta^2 \det(\mathfrak{a},\mathfrak{b}).
\end{aligned}
$$

Therefore $\langle \mathfrak{f}^\perp, y_* \rangle = -\eta^2 \det(\mathfrak{a}, \mathfrak{b})$, and thus

$$\langle \mathfrak{f}, y_* \rangle = \pm(|\mathfrak{f}|^2|y_*|^2 - \langle \mathfrak{f}^\perp, y_* \rangle^2)^{1/2} = \pm(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}.$$

Using that

$$y_* = \frac{\langle \mathfrak{f}, y_* \rangle \mathfrak{f} + \langle \mathfrak{f}^\perp, y_* \rangle \mathfrak{f}^\perp}{|\mathfrak{f}|^2},$$

we deduce that $y_* \in \{y_1, y_{-1}\}$, where for $\varepsilon \in \{-1, 1\}$,

$$y_\varepsilon := \frac{\varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\mathfrak{f} - \eta^2 \det(\mathfrak{a}, \mathfrak{b})\mathfrak{f}^\perp}{|\mathfrak{f}|^2}.$$

Using that $\langle \mathfrak{a}, \mathfrak{f}^\perp \rangle = \alpha \det(\mathfrak{a}, \mathfrak{b})$ and $\langle \mathfrak{b}, \mathfrak{f}^\perp \rangle = \beta \det(\mathfrak{a}, \mathfrak{b})$, we compute that

$$\frac{\alpha + \langle \mathfrak{a}, y_\varepsilon \rangle}{\beta + \langle \mathfrak{b}, y_\varepsilon \rangle} = \frac{(1 - \eta^2 \det(\mathfrak{a}, \mathfrak{b})/|\mathfrak{f}|^2)\alpha + \varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\langle \mathfrak{a}, \mathfrak{f} \rangle}{(1 - \eta^2 \det(\mathfrak{a}, \mathfrak{b})/|\mathfrak{f}|^2)\beta + \varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\langle \mathfrak{b}, \mathfrak{f} \rangle}.$$

Note that the denominator is always positive, since $|\mathfrak{b}||y_\varepsilon| = \eta|\mathfrak{b}| < \beta$. We deduce from $0 < |\mathfrak{f}|^2 = \langle \mathfrak{f}, \mathfrak{f} \rangle = \langle \beta\mathfrak{a} - \alpha\mathfrak{b}, \mathfrak{f} \rangle$ that $\beta\langle \mathfrak{a}, \mathfrak{f} \rangle > \alpha\langle \mathfrak{b}, \mathfrak{f} \rangle$, and thus that

$$\frac{\alpha + \langle \mathfrak{a}, y_1 \rangle}{\beta + \langle \mathfrak{b}, y_1 \rangle} > \frac{\alpha + \langle \mathfrak{a}, y_{-1} \rangle}{\beta + \langle \mathfrak{b}, y_{-1} \rangle},$$

which implies that $y_* = y_1$. □

Proof of Proposition 6.C.2. Recall that

$$\sigma_{P(x)}(e) := \sup_{p \in P(x)} \langle e, p \rangle = \sup_{p \in -D_x c(x, Y)} \langle e, p \rangle.$$

By Proposition 6.C.1, the map $\overline{Y} \to -D_x c(x, \overline{Y})$, $y \mapsto -D_x c(x, y)$ is a continuous bijection, hence

$$\sigma_{P(x)}(e) = \sup_{p \in -D_x c(x, Y)} \langle e, p \rangle = \max_{p \in \partial(-D_x c(x, Y))} \langle e, p \rangle = \max_{y \in \partial Y} -\langle e, D_x c(x, y) \rangle.$$

If $y \in \overline{Y}$, we compute that

$$\begin{aligned}
-D_x c(x, y) &= -D\psi_i(x) D_{\hat{x}_i} \hat{c}(\psi(x), \psi(y)) \\
&= \left(\mathfrak{n}(x) I_2 - \mathfrak{n}(x)^3 x \otimes x, -\mathfrak{n}(x)^3 x\right) \frac{\psi(y) - \langle \psi(x), \psi(y) \rangle \psi(x)}{1 - \kappa\langle \psi(x), \psi(y) \rangle} \\
&= \frac{1}{1 - \kappa\langle \psi(x), \psi(y) \rangle} (\mathfrak{n}(x)\mathfrak{n}(y)y - \mathfrak{n}(x)^2\langle \psi(x), \psi(y) \rangle x - \mathfrak{n}(x)^3\mathfrak{n}(y)(\langle x, y \rangle + 1)x \\
&\qquad + \mathfrak{n}(x)^4(|x|^2 + 1)\langle \psi(x), \psi(y) \rangle x) \\
&= \frac{1}{1 - \kappa\langle \psi(x), \psi(y) \rangle} (\mathfrak{n}(x)\mathfrak{n}(y)y - \mathfrak{n}(x)^2\langle \psi(x), \psi(y) \rangle x - \mathfrak{n}(x)^2\langle \psi(x), \psi(y) \rangle x \\
&\qquad + \mathfrak{n}(x)^2\langle \psi(x), \psi(y) \rangle x) \\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y)y - \mathfrak{n}(x)^2\langle \psi(x), \psi(y) \rangle x}{1 - \kappa\langle \psi(x), \psi(y) \rangle} \\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y)y - \mathfrak{n}(x)^3\mathfrak{n}(y)\langle x, y \rangle x - \mathfrak{n}(x)^3\mathfrak{n}(y)x}{1 - \kappa\mathfrak{n}(x)\mathfrak{n}(y)\langle x, y \rangle - \kappa\mathfrak{n}(x)\mathfrak{n}(y)}.
\end{aligned}$$

If $y \in \partial Y$, then $|y|^2 = \delta_y^{-2} - 1$ and thus $\mathfrak{n}(y) = (|y|^2+1)^{-1/2} = \delta_y$. Therefore

$$\begin{aligned}\sigma_{P(x)}(e) &= \max_{y\in\partial Y} -\langle e, D_x c(x,y)\rangle \\ &= \max_{y\in\partial Y} \frac{\mathfrak{n}(x)\delta_y\langle e,y\rangle - \mathfrak{n}(x)^3\delta_y\langle x,y\rangle\langle e,x\rangle - \mathfrak{n}(x)^3\delta_y\langle e,x\rangle}{1-\kappa\mathfrak{n}(x)\delta_y\langle x,y\rangle - \kappa\mathfrak{n}(x)\delta_y} \\ &= \max_{y\in\partial Y} \frac{\alpha+\langle \mathfrak{a},y\rangle}{\beta+\langle\mathfrak{b},y\rangle} = \max_{|y|=(\delta_y^{-2}-1)^{1/2}} \frac{\alpha+\langle\mathfrak{a},y\rangle}{\beta+\langle\mathfrak{b},y\rangle},\end{aligned}$$

where

$$\begin{aligned}\alpha &:= -\mathfrak{n}(x)^3\delta_y\langle e,x\rangle, & \mathfrak{a} &:= \mathfrak{n}(x)\delta_y e - \mathfrak{n}(x)^3\delta_y\langle e,x\rangle x,\\ \beta &:= 1-\kappa\mathfrak{n}(x)\delta_y, & \mathfrak{b} &:= -\kappa\mathfrak{n}(x)\delta_y x.\end{aligned}$$

We let y_* denote a solution to the above maximum, so that (6.84) holds.

Let $\mathfrak{f} := \beta\mathfrak{a} - \alpha\mathfrak{b}$, so that (6.86) holds. The vector $\mathfrak{f}$ is always nonzero, since otherwise $x = \pm|x|e$ and then, using that $1-\mathfrak{n}(x)^2|x|^2 = \mathfrak{n}(x)^2$ and that, by (6.72), $\mathfrak{n}(x) = \langle\psi(x),\psi(0)\geq\kappa$,

$$\mathfrak{f} = (\mathfrak{n}(x)\delta_y - \kappa\mathfrak{n}(x)^2\delta_y^2 - \mathfrak{n}(x)^3\delta_y|x|^2)e = (\mathfrak{n}(x)^3\delta_y - \kappa\mathfrak{n}(x)^2\delta_y^2)e \neq 0.$$

Since $|x| = (\mathfrak{n}(x)^{-2}-1)^{1/2}$, it always holds that

$$(\delta_y^{-2}-1)^{1/2}|\mathfrak{b}| = \kappa\mathfrak{n}(x)(1-\delta_y^2)^{1/2}|x| = \kappa(1-\mathfrak{n}(x)^2)^{1/2}(1-\delta_y^2)^{1/2} < 1-\kappa\mathfrak{n}(x)\delta_y = \beta.$$

We prove (6.83) by applying Lemma 6.C.3, using that

$$(\delta_y^{-2}-1)\det(\mathfrak{a},\mathfrak{b}) = -\kappa\mathfrak{n}(x)^2(1-\delta_y^2)\det(e,x).$$

This concludes the proof. □

Bibliography

[ABM06] S. Alama, L. Bronsard, and J. A. Montero. "On the Ginzburg-Landau Model of a Superconducting Ball in a Uniform Field". In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 23.2 (2006), pp. 237–267.

[BB00] J.-D. Benamou and Y. Brenier. "A Computational Fluid Mechanics Solution to the Monge-Kantorovich Mass Transfer Problem". In: *Numer. Math.* 84.3 (2000), pp. 375–393.

[BBM21a] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau. *A Linear Finite-Difference Scheme for Approximating Randers Distances on Cartesian Grids*. HAL preprint hal-03125879. 2021.

[BBM21b] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau. "Monotone and Second Order Consistent Scheme for the Two Dimensional Pucci Equation". In: *Numerical Mathematics and Advanced Applications ENUMATH 2019*. Ed. by F. J. Vermolen and C. Vuik. Springer, Cham, 2021, pp. 733–742.

[BBM21c] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau. "Second Order Monotone Finite Differences Discretization of Linear Anisotropic Differential Operators". In: *Math. Comp.* 90.332 (2021), pp. 2671–2703.

[BC97] M. Bardi and I. Capuzzo Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Modern Birkhäuser Classics. Birkhäuser, Basel, 1997.

[BCM16] J.-D. Benamou, F. Collino, and J.-M. Mirebeau. "Monotone and Consistent Discretization of the Monge-Ampère Operator". In: *Math. Comp.* 85.302 (2016), pp. 2743–2775.

[BD19] J.-D. Benamou and V. Duval. "Minimal Convex Extensions and Finite Difference Discretization of the Quadratic Monge-Kantorovich Problem". In: *Eur. J. Appl. Math.* 30.6 (2019), pp. 1041–1078.

[BG15] J. F. Bonnans and S. Gaubert. *Recherche opérationnelle. Aspects mathématiques et applications*. Les Éditions de l'École polytechnique, 2015.

[BJ07] G. Barles and E. R. Jakobsen. "Error Bounds for Monotone Approximation Schemes for Parabolic Hamilton-Jacobi-Bellman Equations". In: *Math. Comp.* 76.260 (2007), pp. 1861–1893.

[BM21] G. Bonnet and J.-M. Mirebeau. *Monotone Discretization of the Monge-Ampère Equation of Optimal Transport*. HAL preprint hal-03255797. 2021.

[BOZ04] J. F. Bonnans, É. Ottenwaelter, and H. Zidani. "A Fast Algorithm for the Two Dimensional HJB Equation of Stochastic Control". In: *ESAIM Math. Model. Numer. Anal.* 38.4 (2004), pp. 723–735.

[BP88] G. Barles and B. Perthame. “Exit Time Problems in Optimal Control and Vanishing Viscosity Method”. In: *SIAM J. Control Optim.* 26.5 (1988), pp. 1133–1148.

[BR98] G. Barles and E. Rouy. “A Strong Comparison Result for the Bellman Equation Arising in Stochastic Exit Time Control Problems and Its Applications”. In: *Comm. Partial Differential Equations* 23.11–12 (1998), pp. 1995–2033.

[Bre91] Y. Brenier. “Polar Factorization and Monotone Rearrangement of Vector-Valued Functions”. In: *Comm. Pure Appl. Math.* 44.4 (1991), pp. 375–417.

[BRS04] D. Bao, C. Robles, and Z. Shen. “Zermelo Navigation on Riemannian Manifolds”. In: *J. Differential Geom.* 66.3 (2004), pp. 377–435.

[BS91] G. Barles and P. E. Souganidis. “Convergence of Approximation Schemes for Fully Nonlinear Second Order Equations”. In: *Asymptotic Anal.* 4.3 (1991), pp. 271–283.

[Car01] M. Carter. *Foundations of Mathematical Economics*. MIT Press, Cambridge, MA, 2001.

[Cas86] E. Casas. “Control of an Elliptic Problem with Pointwise State Constraints”. In: *SIAM J. Control Optim.* 24.6 (1986), pp. 1309–1318.

[Chi+20] L. Chizat et al. “Faster Wasserstein Distance Estimation with the Sinkhorn Divergence”. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Ed. by H. Larochelle et al. 2020.

[CIL92] M. G. Crandall, H. Ishii, and P.-L. Lions. “User’s Guide to Viscosity Solutions of Second Order Partial Differential Equations”. In: *Bull. Amer. Math. Soc.* 27.1 (1992), pp. 1–67.

[CMC16] D. Chen, J.-M. Mirebeau, and L. D. Cohen. “Finsler Geodesics Evolution Model for Region based Active Contours”. In: *Proceedings of the British Machine Vision Conference (BMVC)*. Ed. by R. C. Wilson, E. R. Hancock, and W. A. P. Smith. BMVA Press, 2016, pp. 22.1–22.12.

[CMC17] D. Chen, J.-M. Mirebeau, and L. D. Cohen. “Global Minimum for a Finsler Elastica Minimal Path Approach”. In: *Int. J. Comput. Vis.* 122.3 (2017), pp. 458–483.

[CO08] L. A. Caffarelli and V. I. Oliker. “Weak Solution of One Inverse Problem in Geometric Optics”. In: *J. Math. Sci.* 154.1 (2008), pp. 39–49.

[Coh+18] M. B. Cohen et al. “Solving Directed Laplacian Systems in Nearly-Linear Time through Sparse LU Factorizations”. In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2018, pp. 898–909.

[Cor+04] L. P. Cordella et al. “A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.10 (2004), pp. 1367–1372.

[Cra+20] K. Crane et al. *A Survey of Algorithms for Geodesic Paths and Distances*. arXiv preprint arXiv:2007.10430. 2020.

[CS88] J. H. Conway and N. J. A. Sloane. “Low-Dimensional Lattices. III. Perfect Forms”. In: *Proc. Roy. Soc. London Ser. A* 418.1854 (1988), pp. 43–80.

[CS92] J. H. Conway and N. J. A. Sloane. “Low-Dimensional Lattices. VI. Voronoi Reduction of Three-Dimensional Lattices”. In: *Proc. Roy. Soc. London Ser. A* 436.1896 (1992), pp. 55–68.

[Cut13] M. Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *NIPS’13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Ed. by C. J. C. Burges et al. Curran Associates Inc., Red Hook, NY, 2013, pp. 2292–2300.

[CWL18] Y. Chen, J. W. L. Wan, and J. Lin. "Monotone Mixed Finite Difference Scheme for Monge-Ampère Equation". In: *Journal of Scientific Computing* 76.3 (2018), pp. 1839–1867.

[CWW13] K. Crane, C. Weischedel, and M. Wardetzky. "Geodesics in Heat: A New Approach to Computing Distance Based on Heat Flow". In: *ACM Trans. Graph.* 32.5 (2013), 152:1–152:11.

[Dac08] B. Dacorogna. *Direct Methods in the Calculus of Variations*. Vol. 78. Applied Mathematical Sciences. Springer, New York, 2008.

[Deg+12] P. Degond et al. "An Asymptotic-Preserving Method for Highly Anisotropic Elliptic Equations Based on a Micro-Macro Decomposition". In: *J. Comput. Phys.* 231.7 (2012), pp. 2724–2740.

[Des+21] F. Desquilbet et al. *Single Pass Computation of First Seismic Wave Travel Time in Three Dimensional Heterogeneous Media with General Anisotropy*. HAL preprint hal-03244537. 2021.

[DF14] G. De Philippis and A. Figalli. "The Monge-Ampère Equation and Its Link to Optimal Transportation". In: *Bull. Amer. Math. Soc.* 51.4 (2014), pp. 527–580.

[DJ13] K. Debrabant and E. R. Jakobsen. "Semi-Lagrangian Schemes for Linear and Fully Nonlinear Diffusion Equations". In: *Math. Comp.* 82.283 (2013), pp. 1433–1462.

[DK09] R. Devore and A. Kunoth, eds. *Multiscale, Nonlinear and Adaptive Approximation*. Springer, Berlin, Heidelberg, 2009.

[DMT16] P. M. M. De Castro, Q. Mérigot, and B. Thibert. "Far-Field Reflector Problem and Intersection of Paraboloids". In: *Numer. Math.* 134.2 (2016), pp. 389–411.

[DSV07] M. Dutour Sikirić, A. Schürmann, and F. Vallentin. "Classification of Eight-Dimensional Perfect Forms". In: *Electron. Res. Announc. Amer. Math. Soc.* 13.3 (2007), pp. 21–32.

[DSV12] M. Dutour Sikirić, A. Schürmann, and F. Vallentin. "Inhomogeneous Extreme Forms". In: *Ann. Inst. Fourier (Grenoble)* 62.6 (2012), pp. 2227–2255.

[Dui+18] R. Duits et al. "Optimal Paths for Variants of the 2D and 3D Reeds-Shepp Car with Applications in Image Analysis". In: *J. Math. Imaging Vision* 60.6 (2018), pp. 816–848.

[EG92] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[Erd92] R. Erdahl. "A Cone of Inhomogeneous Second-Order Polynomials". In: *Discrete Comput. Geom.* 8.4 (1992), pp. 387–416.

[FJ17] X. Feng and M. Jensen. "Convergent Semi-Lagrangian Methods for the Monge-Ampère Equation on Unstructured Grids". In: *SIAM J. Numer. Anal.* 55.2 (2017), pp. 691–712.

[FL09] A. Figalli and G. Loeper. "C^1 Regularity of Solutions of the Monge-Ampère Equation for Optimal Transport in Dimension Two". In: *Calc. Var. Partial Differential Equations* 35.4 (2009), pp. 537–550.

[FL21] B. Froese Hamfeldt and J. Lesniewski. *A Convergent Finite Difference Method for Computing Minimal Lagrangian Graphs*. arXiv preprint arXiv:2102.10159. 2021.

[FM14] J. Fehrenbach and J.-M. Mirebeau. "Sparse Non-Negative Stencils for Anisotropic Diffusion". In: *J. Math. Imaging Vision* 49.1 (2014), pp. 123–147.

[FO11] B. D. Froese and A. M. Oberman. "Convergent Finite Difference Solvers for Viscosity Solutions of the Elliptic Monge–Ampère Equation in Dimensions Two and Higher". In: *SIAM Journal on Numerical Analysis* 49.4 (2011), pp. 1692–1714.

[FO13] B. D. Froese and A. M. Oberman. "Convergent Filtered Schemes for the Monge-Ampère Partial Differential Equation". In: *SIAM J. Numer. Anal.* 51.1 (2013), pp. 423–444.

[Fro19] B. Froese Hamfeldt. "Convergence Framework for the Second Boundary Value Problem for the Monge-Ampère Equation". In: *SIAM J. Numer. Anal.* 57.2 (2019), pp. 945–971.

[GH09] C. E. Gutiérrez and Q. Huang. "The Refractor Problem in Reshaping Light Beams". In: *Arch. Ration. Mech. Anal.* 193.2 (2009), pp. 423–443.

[GH14] C. E. Gutiérrez and Q. Huang. "The Near Field Refractor". In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 31.4 (2014), pp. 655–684.

[Gut16] C. E. Gutiérrez. *The Monge-Ampère Equation*. Vol. 89. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, Basel, 2016.

[IL90] H. Ishii and P.-L. Lions. "Viscosity Solutions of Fully Nonlinear Second-Order Elliptic Partial Differential Equations". In: *J. Differential Equations* 83.1 (1990), pp. 26–78.

[KMT19] J. Kitagawa, Q. Mérigot, and B. Thibert. "Convergence of a Newton Algorithm for Semi-Discrete Optimal Transport". In: *J. Eur. Math. Soc.* 21.9 (2019), pp. 2603–2651.

[KO97] S. A. Kochengin and V. I. Oliker. "Determination of Reflector Surfaces from Near-Field Scattering Data". In: *Inverse Problems* 13.2 (1997), pp. 363–373.

[Kom88] H. Komiya. "Elementary Proof for Sion's Minimax Theorem". In: *Kodai Math. J.* 11.1 (1988), pp. 5–7.

[Kry05] N. V. Krylov. "The Rate of Convergence of Finite-Difference Approximations for Bellman Equations with Lipschitz Coefficients". In: *Appl. Math. and Optim.* 52.3 (2005), pp. 365–399.

[Kry87] N. V. Krylov. *Nonlinear Elliptic and Parabolic Equations of Second Order*. Vol. 7. Mathematics and its Applications. Springer Netherlands, 1987.

[KT92] H.-J. Kuo and N.-S. Trudinger. "Discrete Methods for Fully Nonlinear Elliptic Equations". In: *SIAM J. Numer. Anal.* 29.1 (1992), pp. 123–135.

[Lio85] P.-L. Lions. "Two Remarks on Monge-Ampere Equations". In: *Ann. Mat. Pura Appl.* 142.1 (1985), pp. 263–275.

[LN18] W. Li and R. H. Nochetto. "Optimal Pointwise Error Estimates for Two-Scale Methods for the Monge–Ampère Equation". In: *SIAM J. Numer. Anal.* 56.3 (2018), pp. 1915–1941.

[Mar03] J. Martinet. *Perfect Lattices in Euclidean Spaces*. Vol. 327. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Heidelberg, 2003.

[Mir14] J.-M. Mirebeau. "Efficient Fast Marching with Finsler Metrics". In: *Numer. Math.* 126.3 (2014), pp. 515–557.

[Mir16] J.-M. Mirebeau. "Minimal Stencils for Discretizations of Anisotropic PDEs Preserving Causality or the Maximum Principle". In: *SIAM J. Numer. Anal.* 54.3 (2016), pp. 1582–1611.

[Mir18] J.-M. Mirebeau. "Fast-Marching Methods for Curvature Penalized Shortest Paths". In: *J. Math. Imaging Vision* 60.6 (2018), pp. 784–815.

[Mir19] J.-M. Mirebeau. "Riemannian Fast-Marching on Cartesian Grids, Using Voronoi's First Reduction of Quadratic Forms". In: *SIAM J. Numer. Anal.* 57.6 (2019), pp. 2608–2655.

[MTW05] X.-N. Ma, N. S. Trudinger, and X.-J. Wang. "Regularity of Potential Functions of the Optimal Transportation Problem". In: *Arch. Ration. Mech. Anal.* 177.2 (2005), pp. 151–183.

[MW52] T. S. Motzkin and W. Wasow. "On the Approximation of Linear Elliptic Differential Equations by Difference Equations with Positive Coefficients". In: *J. Math. Phys.* 31.1-4 (1952), pp. 253–259.

[NNZ19] R. H. Nochetto, D. Ntogkas, and W. Zhang. "Two-Scale Method for the Monge-Ampère Equation: Pointwise Error Estimates". In: *IMA J. Numer. Anal.* 39.3 (2019), pp. 1085–1109.

[NS04] P. Q. Nguyen and D. Stehlé. "Low-Dimensional Lattice Basis Reduction Revisited". In: *Algorithmic Number Theory*. Ed. by D. Buell. Springer, Berlin, Heidelberg, 2004, pp. 338–357.

[Obe06] A. M. Oberman. "Convergent Difference Schemes for Degenerate Elliptic and Parabolic Equations: Hamilton-Jacobi Equations and Free Boundary Problems". In: *SIAM J. Numer. Anal.* 44.2 (2006), pp. 879–895.

[Obe07] A. Oberman. "The Convex Envelope is the Solution of a Nonlinear Obstacle Problem". In: *Proc. Amer. Math. Soc.* 135.6 (2007), pp. 1689–1694.

[Obe08] A. M. Oberman. "Wide Stencil Finite Difference Schemes for the Elliptic Monge-Ampère Equation and Functions of the Eigenvalues of the Hessian". In: *Discrete Contin. Dyn. Sys. Ser. B* 10.1 (2008), pp. 221–238.

[OS09] S. Ohta and K.-T. Sturm. "Heat Flow on Finsler Manifolds". In: *Comm. Pure Appl. Math.* 62.10 (2009), pp. 1386–1433.

[Par+19] S. Parisotto et al. "Anisotropic Osmosis Filtering for Shadow Removal in Images". In: *Inverse Problems* 35.5 (2019), p. 054001.

[Ran41] G. Randers. "On an Asymmetrical Metric in the Four-Space of General Relativity". In: *Phys. Rev.* 59.2 (1941), pp. 195–199.

[Sch09a] A. Schürmann. *Computational Geometry of Positive Definite Quadratic Forms. Polyhedral Reduction Theories, Algorithms, and Applications*. University Lecture Series. American Mathematical Society, Providence, RI, 2009.

[Sch09b] A. Schürmann. "Enumerating Perfect Forms". In: *Contemp. Math.* 493 (2009), p. 359.

[Sel74] E. Selling. "Uber die Binären und Ternären Quadratischen Formen". In: *J. Reine Angew. Math.* 77 (1874), pp. 143–229.

[Sin64] R. Sinkhorn. "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices". In: *Ann. Math. Stat.* 35.2 (1964), pp. 876–879.

[Sol+15] J. Solomon et al. "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains". In: *ACM Trans. Graph.* 34.4 (2015), 66:1–66:11.

[Tru14] N. S. Trudinger. "On the Local Theory of Prescribed Jacobian Equations". In: *Discrete Contin. Dyn. Syst.* 34.4 (2014), pp. 1663–1681.

[Var67] S. R. S. Varadhan. "On the Behavior of the Fundamental Solution of the Heat Equation With Variable Coefficients". In: *Comm. Pure Appl. Math.* 20.2 (1967), pp. 431–455.

[Vil03] C. Villani. *Topics in Optimal Transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003.

[Vil09] C. Villani. *Optimal Transport. Old and New*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[Vor08] G. Voronoi. "Sur quelques propriétés des formes quadratiques positives parfaites". In: *J. reine angew. Math.* 133 (1908), pp. 97–178.

[Wei98] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart, 1998.

[Yan+18] F. Yang et al. "Geodesic via Asymmetric Heat Diffusion Based on Finsler Metric". In: *Computer Vision – AACV 2018*. Ed. by C. V. Jawahar et al. Springer, Cham, 2018, pp. 371–386.

[YC16] F. Yang and L. D. Cohen. "Geodesic Distance and Curves through Isotropic and Anisotropic Heat Equations on Images and Surfaces". In: *J. of Math. Imaging Vision* 55.2 (2016), pp. 210–228.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Discrétisation aux différences finies monotones d'équations aux dérivées partielles dégénérées elliptiques en utilisant la première réduction de Voronoi

Mots clés : Différences finies monotones, ellipticité dégénérée, première réduction de Voronoi, distances de Randers, équation de Pucci, équation de Monge-Ampère

Résumé : Dans cette thèse, nous montrons comment la première réduction de Voronoi permet de construire des discrétisations aux différences finies monotones sur grilles cartésiennes de certains opérateurs différentiels dégénérés elliptiques. Nous recommandons une discrétisation particulière, consistante à l'ordre deux, d'opérateurs différentiels linéaires anisotropes en dimensions deux et trois comprenant à la fois des termes d'ordres un et deux. Nous prouvons la quasi-optimalité de cette construction. Nous étudions certaines propriétés de régularité et de compacité de la première réduction de Voronoi en dimension quatre. Nous concevons une méthode permettant d'approcher efficacement des distances de Randers et des distances de transport optimal associées, en utilisant un principe de grandes déviations. Nous discrétisons les opérateurs de Pucci et de Monge-Ampère. Les discrétisations obtenues s'écrivent comme des maxima d'opérateurs discrets ; en dimension deux, nous montrons que ces maxima admettent des expressions de forme fermée, ce qui réduit le coût numérique de leur évaluation. Nous étudions le caractère bien posé et, dans certains cas, la convergence d'un schéma numérique pour le second problème aux limites pour l'équation de Monge-Ampère. Nous présentons une application numérique au problème du réfracteur en champ lointain en optique non imageante.

Title: Monotone finite difference discretization of degenerate elliptic partial differential equations using Voronoi's first reduction

Keywords: Monotone finite differences, degenerate ellipticity, Voronoi's first reduction, Randers distances, Pucci equation, Monge-Ampère equation

Abstract: In this thesis, we show how Voronoi's first reduction may be used in order to build monotone finite difference discretizations on Cartesian grids of some degenerate elliptic differential operators. We recommend a specific, second-order consistent discretization of two- and three-dimensional linear anisotropic differential operators involving both a first- and a second-order term. We prove the quasi-optimality of this construction. We study some properties on the regularity and the compactness of Voronoi's first reduction in dimension four. We design a method allowing to efficiently approximate Randers distances and associated optimal transport distances, using a large deviations principle. We discretize the Pucci and Monge-Ampère operators. The resulting discretizations are written as maxima of discrete operators; in dimension two, we show that these maxima admit closed-form formulae, reducing the numerical cost of their evaluation. We study the well-posedness, and in some cases the convergence, of a numerical scheme for the second boundary value problem for the Monge-Ampère equation. We present a numerical application to the far-field refractor problem in nonimaging optics.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France